

**TOWARDS AUTOMATED AXIOM GENERATION: A
SEMI-AUTOMATED APPROACH TO GENERATING
“KNOWLEDGE AND RULE BASE” CORPORA FROM TEXT
NARRATIVES**

Anirudh Prabhu

Submitted in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Approved by:
Peter Fox, Chair
Deborah McGuinness, Co-Chair
James Hendler
Sergei Nirenburg
Xiaogang Ma



Department of Multidisciplinary Science
Rensselaer Polytechnic Institute
Troy, New York

[August 2021]
Submitted July 2021

© Copyright 2021
by
Anirudh Prabhu
All Rights Reserved

CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
ACKNOWLEDGMENT	xi
ABSTRACT	xiii
1. INTRODUCTION	1
1.1 Thinking Machines	1
1.2 Artificial Intelligence and a Big Divide	1
1.2.1 Human Input in Artificial Intelligence	2
1.3 Rules in GOFAI	2
1.4 Problem Statement	3
1.5 Proposed Research Methodology and Contributions	3
1.6 Impact	6
1.7 Dissertation Summary	6
2. BACKGROUND AND RELATED WORK	7
2.1 Artificial Intelligence	7
2.1.1 History	7
2.2 Rule-Based AI	7
2.2.1 History	7
2.2.2 Components	8
2.2.3 System Architecture	8
2.2.4 Inference Techniques	9
2.2.5 Advantages and Disadvantages	9
2.2.5.1 Advantages	9
2.2.5.2 Disadvantages	10
2.3 Situation Calculus	10
2.3.1 Different Chains of Thought	10
2.3.2 Components	10
2.3.3 Axioms	11
2.3.4 The Frame Problem in Situation Calculus	12
2.3.4.1 Reiter's Solution to Frame Problem	12
2.3.5 A Common Use Case in Situation Calculus	12

2.3.5.1	GOLOG	13
2.4	Deep Learning	14
2.4.1	History: The Birth, Decline and Resurgence of Neural Networks . . .	14
2.4.2	Recurrent Neural Networks	15
2.4.2.1	Long Short Term Memory	15
2.5	Using Sentence Ordering Algorithms to Arrange a Text Narrative in the Form of a Fact Base	15
2.5.1	Why is It Important in a Fact Base	16
2.5.2	State of the Art	16
2.5.3	Method	18
2.5.3.1	Topical Clustering + Improvements	18
2.5.3.2	Neural Network Approach	19
2.5.3.3	Recurrent Neural Networks + Long Short Term Memory . .	20
2.5.4	Sentence Ordering Evaluation	21
2.5.5	Results	22
2.5.6	Conclusion and Future Work	22
2.5.6.1	Capsule Networks	23
2.6	Information Extraction	24
2.6.1	Types of Information Extraction	24
2.6.2	Event Extraction	24
2.6.3	Fact Extraction	25
3.	CREATING RULES IN TEXT FORM BASED ON A NARRATIVE USING CROWD- SOURCING	26
3.1	Introduction	26
3.2	Crowdsourcing Dataset Creation	26
3.3	Background	27
3.3.1	Data	29
3.3.2	Limitations	29
3.3.2.1	Solutions	29
3.4	Method	30
3.5	Incentive	31
3.5.1	Incentive Adjustment	31
3.6	Curation and Evaluation	32
3.7	Potential Issues	34
3.8	Results and Conclusions	34

4.	OBSERVATIONS ON AUTOMATED AXIOM GENERATION WHEN CROWD-SOURCING IS AN INADEQUATE PROXY FOR HUMAN EXPERT IN THE LOOP	36
4.1	Background	36
4.2	Observations (Lessons Learnt)	36
4.2.1	Rules from the Crowdsourcing Experiment are Inconsistent	37
4.2.2	Curation Takes a Long Time	37
4.2.3	Axiom Boundaries are Difficult to Frame	38
4.2.4	Rule Extraction Has High Complexity	38
4.2.5	Considered Choice of Fictional Stories and Its Potential Limitations	39
4.3	Human Evaluation of Crowdsourcing Results	39
4.3.1	Evaluation Criteria	40
4.3.1.1	Level 1: Acceptable Criteria for Rules	40
4.3.1.2	Level 2:Fixing Grammatical Errors and Sentence Construction Issues	41
4.3.1.3	Level 3:Labeling the Commonsense Rules	41
4.3.2	Observations	42
4.3.2.1	Overview	42
4.3.2.2	Individual Rule Observations and Limitations	42
4.4	Crowdsourcing in the Context of Expert Replacements	44
4.4.1	Boundaries of Crowdsourcing Tasks and Their Success	46
4.5	The "Humans in the Loop" Discussion and "Experts in the Loop" Distinction	46
4.5.1	Citizen Science Explorations and Their Boundaries	48
4.6	Conclusion	49
5.	CONVERTING TEXT RESULTS OF THE CROWDSOURCING EXPERIMENT INTO FORMAL RULES AXIOMS	53
5.1	Background	53
5.2	Algorithm/Methodology	53
5.2.1	NLP Based Methods	54
5.2.1.1	How to Handle Rules	54
5.2.1.2	Complex Axioms	55
5.2.1.3	Simple Axioms	55
5.3	Results	56
5.4	Observations and Limitations	56
5.5	Future Work	58

5.5.1	Machine Translation	59
5.5.1.1	Deep Learning	59
5.5.1.2	Evaluation	60
6.	EVALUATING RULES EXTRACTED FROM A TEXT NARRATIVE	66
6.1	Need for Evaluation	66
6.2	What Can We Learn from Existing Evaluation Approaches?	66
6.2.1	Ontology Evaluation	66
6.2.2	Expert System Evaluation	68
6.3	Are They Directly Applicable?	68
6.3.1	Adapting Metrics	69
6.4	New Metric - Coverage	70
6.5	Results	71
6.5.1	Interpretability	72
6.5.2	Coverage	72
6.5.3	Lawfulness	73
6.6	Future Work	73
6.6.1	Accuracy	73
6.7	Testing Wider Applicability	74
7.	CONCLUSIONS AND FUTURE WORK	75
7.1	Thesis Review	75
7.2	Lessons Learnt	76
7.3	Impact and Future Applications	77
7.4	Future Work	78
7.4.1	Short Term	78
7.4.2	Mid Term	79
7.4.3	Long Term	79
	REFERENCES	81
	APPENDICES	96
	A. IRB APPROVAL FOR CROWDSOURCING EXPERIMENT	96
	B. CROWDSOURCING EXPERIMENT IN MECHANICAL TURK	101
	C. CROWDSOURCING RULES EXAMPLES FROM PUBLISHED BATCHES IN MTURK	105

D. RESULTS FROM CROWDSOURCING EVALUATION	108
E. TEXT TO RULE CONVERSION FULL RESULTS	110
F. EVALUATION RESULTS FOR THE SAAG WORKFLOW	111
F.1 Interpretability, Coverage and Normalized Coverage	111
F.2 Lawfulness	111

LIST OF TABLES

2.1	Results from the LSTM sentence ordering model.	22
D.1	Data description for crowdsourcing evaluation.	108
D.2	Data preview for crowdsourcing evaluation.	109
E.1	Data description for text to rules conversion results.	110
E.2	Data preview for text to rules conversion results.	110
F.1	Data description for evaluation results for the SAAG workflow.	111
F.2	Data preview for evaluation results for the SAAG workflow.	111

LIST OF FIGURES

1.1	Proposed research methodology. The column in the left presents a high level overview of the proposed method, whereas the middle column presented a detailed workflow. The solid blocks and solid lines indicate the main workflow presented in this thesis. Dash-lines and boxes represent the techniques used to complete a specific step in the main workflow.	4
2.1	Situation calculus workflow. The user starts with an initial state (Situation S_0), and foundational axioms that hold true throughout the narrative. Next, we get a sequence of actions that occur in the narrative, and the fact base comprises mainly of these actions. Lastly, we document axioms/rules based on what we understand about the domain (world presented in the narrative).	13
4.1	A mosaic plot representing the properties of the rules that have undergone the evaluation process as described in Section 4.3.	50
4.2	Mosaic plot representing the properties of the rules that have undergone the evaluation process as described in Section 4.3	51
4.3	Comparison of the crowdsourcing tasks. The X-axis shows complexity and the Y-axis shows the likelihood of the results being used in an automated workflow.	52
5.1	Our proposed method for converting results of the crowdsourcing experiment (rules in text form) into formal rules. In the figure, the solid lines and shapes represent the main workflow, and the dotted lines represents comments or notes needed at the corresponding step for easily understanding the method.	62
5.2	We identify the "type" of axiom, based on the results of the crowdsourcing evaluation. The evaluators look for the presence of an antecedent and consequent. A rule with multiple subject object pairs, or a antecedent and consequent is classified as a "complex axiom" by our algorithm. If it meets neither of the 2 criteria, the axiom is considered "simple".	63
5.3	Processing complex axioms involves separating out the antecedent and consequent, identifying the concepts mentioned in those parts, and storing those concepts in the rule base.	64
5.4	Processing simple axioms broadly involves identifying the subject, action(verb) and object. We extract triples using the prepositions identified by the python implementation of 'clauseIE'.	65
B.1	Examples of the batches published in Amazon Mechanical Turk. Each batch summary shows the average time taken per task. Number of tasks (assignments), the creation and completion time of each task.	101
B.2	Examples of the results in Amazon Mechanical Turk. The figure shows the interface presented to requester for approving or rejecting rules.	102

B.3	Screenshots from the Amazon Mechanical Turk interface that are presented to the participants of the crowdsourcing experiment.	103
B.4	Instructions from the Amazon Mechanical Turk interface that are presented to the participants of the crowdsourcing experiment.	104
F.1	Detailed output loading all rule bases into the SWI-Prolog environment.	111

ACKNOWLEDGMENT

It has been a long road to in my journey to complete my dissertation. Without the guidance, support, patience, tolerance and encouragement of those around me, I would not have been able to see it through to the end.

To Peter Fox, my advisor and my mentor. You were the one who told me I should be pursue a PhD when I first met you as a masters student. I miss you Peter and will forever be grateful to you for seeing potential in me (when I didnt see any in myself) and making me act on it. Your influence on my thoughts and career will stay with for the rest of my life. My PhD journey would not have as exciting, entertaining, and eye-opening without your guidance. I have learnt just as much observing you in action as I have from our discussions. Your belief in the value of my thesis work and the importance of seeing it through is what kept me going during the times I was doubting myself. Studying under and working with you has made me strive to be a better researcher, a better collaborator and a better person. My only regret is not getting to celebrate my PhD completion with you. I look forward to completing all the ideas that we had to leave unfinished.

To Deborah McGuinness, thank you for guiding me through a very tough last few months and helping me see this dissertation to its finish. Your ontology engineering class was the first time we met and since then you constantly provided me advice whenever I approached you with questions. You agreed to be on my committee and later my advisor and committee chair after Peters passing. Thank you for all your help and support.

To Kathy Fontaine, thank you for all the help and support all these years and especially the last few months. You made time for me and listened to my worries, thesis related or otherwise. I am really thankful for all your guidance and for you building my confidence. Without your help, completing my dissertation would not have been nearly as stress free and smooth. Thank you!

To Jim Hendler, Sergei Nirenburg and Xiaogang (Marshall) Ma, thank you for accepting to be on my committee, for your guidance and direction in shaping my thesis and help making it better.

The Tetherless World Constellation (TWC) has been my home for the last 7 years. My scientific career began here, and I could not have asked for a better setting to get the right scientific training and ample opportunities and freedom to apply my skills in the real

world. Multidisciplinary isn't just a word but a way of life at Tetherless World, and that says everything about my decision to pursue a PhD in Multidisciplinary Science. I would like to thank all the staff at TWC, both past and present for making my time at Tetherless great! I especially want to thank Jacky Carley and Melissa Anderson. Jacky, you were the glue that held TWC together and kept everything running smoothly. Hope you have a great retired life! Melissa, thank you for making sure I got paid on time and I could focus on my work worry-free.

To my colleagues and friends, both from TWC and outside, it is hard to articulate how much fun it has been working and spending time with you all. I am apprehensive of putting a list of names just because I am scared I will miss somebody. But know that I cherish your friendship and thank you for tolerating me all these years. Could not have asked for a better group of folks to enjoy restaurant visits, hikes, karaoke, game sessions, movie nights or just generally hanging out.

To my collaborators, thank you for giving me the opportunity to work with you all. Working with you all expanded my horizons and taught me how a good scientific collaboration works. I look forward to our continued collaboration and I am sure I will continue to learn a lot from all of you.

To my family, a thank you just doesn't seem enough. My parents shaped my outlook on life and that has helped me overcome all the hurdles thrown at me. The almost daily conversations with my dad all these years have helped me keep my sanity through tough times, especially during covid and the loss of Peter. Dad, your unconditional support and encouragement mean a whole lot more to me than I have been able to express. Miss you Mom, love you Dad.

ABSTRACT

With the exponential rise of data in recent years, deep learning has risen to be one of the most prominent forms of artificial intelligence. With many successful applications, deep learning has helped researchers build machines that successfully complete human tasks previously thought to be very difficult. For example, restoring color to black and white photos, image captioning, voice generation, restoring sound in silent videos, lip reading from videos etc. are some very interesting applications being explored and with deep learning. Even with success in a breadth of applications, there are still problems that deep learning has not been able to solve. For example, scalability, understanding context, or examining and understanding the inner workings of deep neural networks themselves remain unsolved problems. The crux of the deep learning approach are layers (input, hidden and output). These layers are adjustable to a given corpus and mostly opaque to interpretation or explanation. Current approaches to Artificial Intelligence/Machine Learning rely on an entire corpus, i.e. they use the entire content with noise, bias, etc. These approaches have achieved high success across fields like computer vision, natural language processing, image captioning etc., require a very large amount of training data to accurately understanding the mappings between the input and output embeddings for the deep learning experiment. In this thesis, we ask the question, "What if, intelligent information extraction (both entity and relation) were able to provide a curated corpus for deep network learning?" Curation in this context, addresses eliminating all the "non-essential" parts of text, and simply focusing on the actions, agents and events involved in a text corpus, and the rules that highlight the effects of these actions and change in the narrative.

What is needed to achieve this task, is the ability to recognize key entities and map the situational changes occurring in the corpus to specific triggers (such as actions or events), like those seen in axioms in a rule base. Automated axiom creation is a difficult research problem to solve. Most of the work in this area focuses on rules extracting rules from text that explicitly mentions the rule in text. In most old fashioned AI systems, rules are developed with an understanding of the domain and reading between the lines where required to see what action/events could trigger a particular response in the narrative. An automated axiom creation method that completes such a task is still an unexplored and unsolved problem, and the focus of this thesis.

The biggest hurdle in exploration of this research problem, is the availability of data (or the lack thereof) where implicit rules are documented for a text narrative. In this thesis, we have developed a novel semi-automated method to generate axioms/rules for a set of text narratives, using crowdsourcing and known natural language processing techniques. We begin with textual narrative such as those in novels, computer manuals, but also in view are scientific works. We then document rules for the given narrative by using Amazon Mechanical Turk, a crowdsourcing platform known to aid in the creation of high-quality datasets. We have found that the usage of a crowdsourcing platform works well for narratives that do not require any expertise, like those seen in novels, and are able to provide textual rules for the narrative which may not be explicitly stated in the text. The next step would be to process these narratives and their rules into knowledge bases and rule bases, where the key concepts and relationships need to be extracted from both the knowledge bases (in the form of triples) and rule base. We have also developed an approach to converting the results of the crowdsourcing experiment (rules in text form), to formal rules in a rule base. These are developed based on known NLP information extraction techniques, like POS tagging, co-reference resolution etc. The overall goal of this thesis is a novel method to extract key information and rules from a narrative , in order to create a set of knowledge bases and rule bases. After examining the results of the crowdsourcing experiment, we found a set of boundaries for the usability of crowdsourcing as tool or means to overcome the automation bottleneck. We also discuss the required distinction of the terms "Humans in the loop" and "Experts in the loop", and provide a platform for fleshing out the framework for experts in the loop for scientific workflows. Finally we also developed a method to evaluate "knowledge base - rule base" corpora for any logical language in any domain.

To construct such a "situational narrative", a formalism such as situation calculus stands out as an obvious choice for knowledge representation but is heretofore an unexplored option in explaining what is going on in deep learning. At the heart of such a capability may be a learned formalization of the situation and perhaps even the identification of changes in state or fluent(s) (situation) over iterations or after learning interactions.

CHAPTER 1

INTRODUCTION

1.1 Thinking Machines

”Creating a machine that can think” or ”exhibit intelligent behaviour” has always been one of the goals of humankind [1]–[5]. Historically, the idea of thinking machines have captivated audiences. Broadly, our interest in creating ”thinking machines” comes from 2 main points of view :

- Creating machines that could perform complex and difficult tasks for humans, sometimes better than humans [6],[7].
- Understanding how machines ”understand the data” and ”make decisions” while performing these tasks [8],[9].

These 2 views go hand in hand with research in the field of artificial intelligence. AI research has made progress by having machines successfully perform tasks that were thought to be purely ”human” and impossible for ”machines” at the time. From the checkers programs of the early 50s [10], to ELIZA (the first Chatbot) in the 60s [11], to MYCIN in the 70s which identified bacteria that caused severe infection [12], to Deepblue in the 90s to Watson [13] and AlphaGo in the 21st century [14], artificial intelligence research has made a splash around the world with its successful applications and has continued to capture people’s imaginations.

1.2 Artificial Intelligence and a Big Divide

Artificial Intelligence research can broadly be segregated into 2 approaches:

- **Reasoning-based/Symbolic AI** : Also called ”Good Old Fashioned AI” (GOFAI), this approach focuses on Thomas Hobbes’ idea that ”ratiocination is computation” [15],[16]. As stated in the name, this approach normally follows the open-world assumption (with some exceptions), where ”what is not known to be true or false is interpreted as unknown information, not as negative information” [17]. GOFAI normally use axioms or production rules to make inferences in order to answer the questions posed. These rules are input by an expert in the AI system’s application domain.

- **Machine Learning** : "Machine learning can be broadly defined as computational methods using past data to improve performance or to make accurate predictions" [18]. Machine learning applications follow the closed world assumption, where currently unknown information is interpreted as false. Machine Learning applications have been at the forefront of AI research in recent years. Machine Learning techniques can be supervised, unsupervised or reinforced. Deep Learning is a more specific sub-field of machine learning focusing on the application of specialized neural networks [19].

For a more detailed history of the 2 approaches, refer to Chapter 2 of this thesis. On studying and understanding the evolution of the field of Artificial intelligence, some important gaps appear. One of most interesting aspects of Artificial intelligence research is that, until very recently, the 2 approaches of AI were studied and improved in parallel, but rarely were the approaches combined to leverage each of their advantages.

1.2.1 Human Input in Artificial Intelligence

One of the goals of artificial intelligence has been to teach computers to perform tasks that humans can perform, with relative ease. Over time, researchers have achieved success in having machines perform "human" tasks of varying difficulty. Machine Learning based applications have been known to train on existing data (supervised) or learn to perform a task from scratch (unsupervised) in order to achieve their desired goal. So there is minimal human interference in the implementation of the AI system. Reasoning-based AI applications (like expert systems) though, which were at the peak of their popularity in the 1980's, require human (expert) input in order to perform the assigned task. Expert systems use a rule base in order to reason about a specific domain. The expert system takes in facts and uses "human-input" rules to infer results from the rule-base.

1.3 Rules in GOFAI

One of the essential steps in using old fashioned AI is to successfully infer/predict future scenarios by writing the axioms for the domain. Writing the axioms or rules for a domain is a task that has to be performed by humans, and in most cases experts in the domain. One of the biggest drawbacks of expert systems is the time and manpower it takes to "maintain" these expert systems as it scales up. This maintenance mainly refers to the iterative updating of these axioms [20]. Historically, scaling up experts systems became

increasingly difficult, as the data size began to rise or the scope began to broaden [20],[21]. This is one of the important factors that led to the downfall of mainstream large scale expert systems and brought with it the "second AI winter" [1],[21].

Thus it important to understand and if possible, automate the axiom/rule creation process for Reasoning based AI.

1.4 Problem Statement

Semi-automated/Automated axiom creation is a difficult research problem to solve. Most of the work in this area focuses on extracting rules from text documents with formal language, that explicitly mention rules in the text [22]–[29]. So the main question that inspired this thesis is, **"To what extent can we fully automate the axiom generation process for GOFAI?"**. In most cases where rule bases and knowledge bases have to be built, the rules are developed with an understanding of the domain and reading between the lines where required, to see what actions/events would trigger a particular response in the narrative. This is why writing rules/axioms has always required human (expert) input. To build a completely automated axiom generation algorithm, the machine must be able to pick the key concepts from the text and state how the occurrence of an action or event triggers the next action or set of actions in the text. The machine also must identify fluents or conditions that are not stated explicitly in the text, but would be clear to human readers. This has not been achieved in any of the automated rule extraction research so far and remains a challenging problem in AI. In our quest to solve the above problem, the biggest hurdle has been the availability of data, particularly where implicit rules are documented for a text narrative. In this thesis, we develop a semi-automated method to generate axioms/rules for a set of text narratives, using known NLP techniques and crowdsourcing. The dataset generated by this thesis may be used to train a completely automated axiom/rule generation model.

1.5 Proposed Research Methodology and Contributions

We have chosen short stories from Project Gutenberg¹ [30] as the main dataset for executing our workflow. The first (optional) step in the Semi-Automated Axiom Generation (Hereafter called SAAG) workflow (Figure 1.1) is ordering sentences from these stories into

¹<https://www.kaggle.com/shubchat/1002-short-stories-from-project-guttenberg>

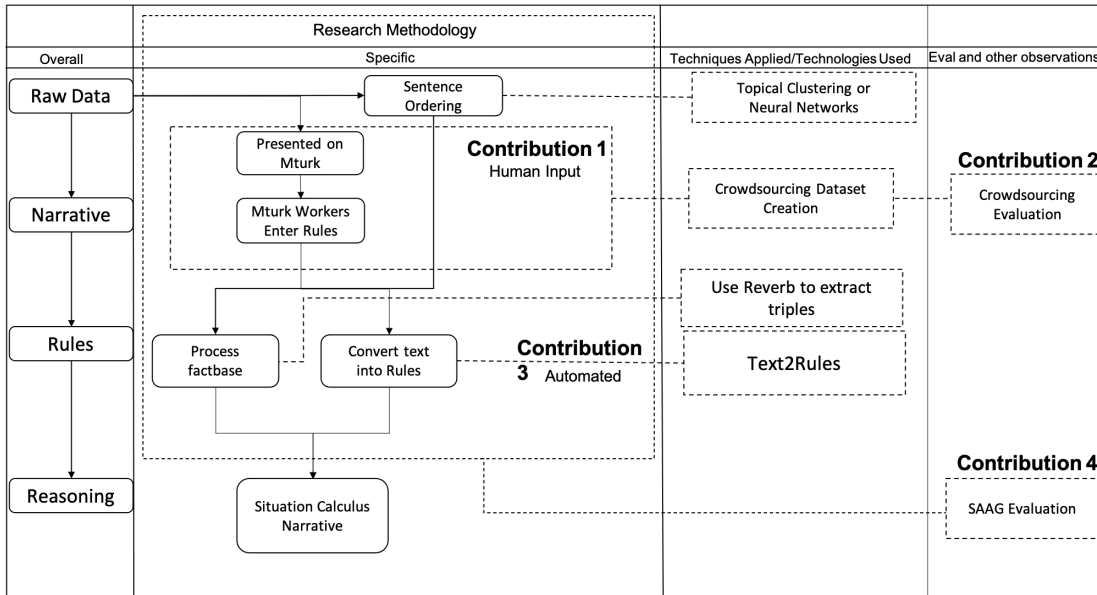


Figure 1.1: Proposed research methodology. The column in the left presents a high level overview of the proposed method, whereas the middle column presented a detailed workflow. The solid blocks and solid lines indicate the main workflow presented in this thesis. Dash-lines and boxes represent the techniques used to complete a specific step in the main workflow.

the order of those commonly seen in fact bases/knowledge bases. This step is more important for certain GOFAI languages and formalisms over others. For example, Situation Calculus, a logical language for representing dynamic changes, has a need for sentence ordering [31], the order in which the events are represented in the fact base matter in the reasoning and inferences made by the machine. For knowledge bases in ontologies, it is important to present facts about a single subject together for easy human readability, but not needed for machine to make inferences. Pronouns and references are fairly common in text addressing multiple subjects like a story. Stories also often jump from topic to topic or from one timeline to another. While this makes for an interesting read, it is not the desired output for a fact base. Hence we recommend using a combination of topical clustering, NLP techniques and/or deep neural networks to predict a sentence order that would best suit the narrative.

The first step of the SAAG workflow is a crowd-sourced approach to document rules about a domain, i.e. the world in the stories, in text form. This step occurs in parallel to the sentence ordering. Here we use a crowdsourcing service, like amazon mechanical turk², and ask people to read a story presented to them and write rules they see obeyed in the

²<https://www.mturk.com/>

story. The rule output by the participants may be a conditional rule or a general axiom. The participants in the data collection experiment are paid based on the specific task assigned to them.

After examining the results of the crowdsourcing experiment, a few interesting points came to light. There were inconsistencies in the rules put out by the participants of the crowdsourcing experiment. We examined these inconsistencies, elaborated on their causes and how to fix them, so that the rules may be useful in creating situation calculus narratives. Additionally, after examining the history of crowdsourcing, its uses, successes and failures along our axiom generation workflow, we assert that *"As the complexity of the tasks increase, crowdsourcing results become unreliable for use in automated workflows"*. Exploring the inconsistencies in the crowdsourcing results and its use in axiom generation, we showed the boundaries of crowdsourcing as an approach learned that *"Crowdsourcing is an inadequate proxy for human experts"*. This is the **first contribution** of the dissertation.

By examining the crowdsourcing in the context of expert replacements and the history of "humans in the loop" workflows, we showed the need for the term "expert in the loop" to be distinct from "humans in the loop", because there are distinct considerations and factors to include experts in the loop vs non-experts in the loop. This is the **second contribution** of the thesis. By starting this discussion, we provide a platform for fleshing out the framework for experts in the loop for scientific workflows.

The next step of the SAAG workflow involves processing the results obtained from the crowdsourcing experiment and converting them into formal rules. We developed a semi-automated approach using NLP based information extraction to achieve this goal. As part of this work, we explore the boundaries of this approach, document the limitations of python implementations of known NLP techniques and suggest potential improvements and future work for this field. With this step of the SAAG workflow, we showed that NLP-based information extraction methods are limited when the boundaries of patterns to be extracted are difficult to define. This is the **third contribution** of the thesis.

For the final step we developed a method to evaluate the rules generated from the execution of the SAAG workflow. Thus, the **fourth contribution** of this thesis is a multi-component evaluation method for rules generated for a knowledge base in any domain and any logical language. This method assesses whether the generated rules can be used to reason over the knowledge base, and how the content of the rules affects the reasoning and

consequently, the predictions or answering capabilities of the situation calculus system. We formulate and apply a set of metrics used to evaluate the various components (Concepts, Actions, Conditions) of the rules individually or combined.

1.6 Impact

As noted in Section 1.4, most of the approaches in automated or semi-automated rule extraction from text use datasets where the rules or conditions for the rule have to be explicitly stated in the text. This makes it difficult to gain rules from text that are not written using a structured template, which includes stories and many other text data sources. This thesis takes a step towards the goal of identifying rules and patterns in text that are not explicitly written. The results of the contributions presented in this thesis may be used to achieve the above stated goal in a fully automated manner, using machine learning methods. The results of the crowdsourcing experiment may be used to do the same, but for a text description. Both of the above goals may be solved by applying abstractive summarization techniques [32],[33] on the knowledge and rule bases. Since that is out of scope for this thesis, we will consider it future work for the presented thesis. More details on the impact and potential future applications of this work can be found in section 7.3.

1.7 Dissertation Summary

Chapter 2 explains the history and related work in the three research areas that led to the conceptualization of the idea for this thesis and explorations into the optional sentence ordering step of the SAAG workflow. Chapters 3,4 and 5 cover the three main contributions that enabled the execution of the proposed methodology (see Figure 1.1). Chapter 6 describes the metrics/method used to evaluate the results of the proposed methods (both the individual contributions and the overall pipeline). Chapter 7 describes the conclusions drawn and lessons learnt from the experiments and execution of the methods described in chapters 3,4 and 5, and planned future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Artificial Intelligence

2.1.1 History

The birth of artificial intelligence as a field of research can be traced back to 1956. John McCarthy, a professor of mathematics at Dartmouth College organized a 2 month workshop to consolidate and develop ideas for the ideal thinking machine. He named this field "Artificial Intelligence" [3].

2.2 Rule-Based AI

2.2.1 History

The first successful set of applications in AI seen in the 1970s and 80s were rule based systems. A rule-based system typically consists of 'if-then' rules, which can be used to answer complex questions or for predictive decision making. The rules in such a system could be written by human experts or could be based on available data [34].

As rule based systems gained popularity in the 1980s, stand alone expert systems were seen as the norm in AI research. But soon, this field of research ran into some hurdles. One of the major challenges of rule-based AI is the generation of the rules itself. Real world systems required completion of complex tasks, which in turn required a larger rule base. This proved to be the undoing of many of the rule-based systems in 1980s. One of the earliest commercially successful rule based expert system was 'R1' [35]. This rule-based production system was used to configure VAX-11/780 computer systems. "Given a customer's order, R1 determined what modifications had to be made to the order for reasons of system functionality and produced a number of diagrams showing how the various components on the order are to be associated" [35]. As the popularity and usage of R1 increased over time, there was a need for addition and modification of rules to make the system work efficiently. In four years the number of rules went from approximately 780 to 3250 [20]. By the 1990s, the system which replaced 75 people, actually needed 150 people to keep it running [21]. Most rule-based systems ran into the similar issues and thus either discontinued or used in a limited domain environment.

Other examples of popular rule based expert systems are:

- MYCIN: MYCIN was a rule based "AI program designed to (a) provide expert level solutions to complex problems, (b) be understandable and (c) be flexible enough to accommodate new knowledge easily" [36].
- PROSPECTOR: The PROSPECTOR system helped geologists in exploring for hard rock mineral deposits. It was intended to emulate the reasoning of an experienced exploration geologist in assessing the likelihood that a given prospect site or region contained an ore deposit of a specific type [37].
- DENDRAL: DENDRAL (which stands for DENDRitic ALgorithm) was the first rule based system applied to solve a "real world" problems. The algorithm defines a set of possible solutions through which the program can search for likely solutions [38].
- HSPEXP: HSPEXP was a system "created to assist less experienced modelers of calibration of a watershed model and to facilitate the interaction between the modeler and the modeling process not provided by mathematical optimization" [39].

2.2.2 Components

Every rule-based expert system consists of the following components [21],[40]:

- Facts: Express assertions about properties, relations, propositions and generally anything relevant to the beginning state of the system.
- Rules: Consist of all the actions that should be taken within a particular scope to specify how to act on the assertion set. Rules express a conditional with an antecedent and a consequent and follows an 'if-then' form.
- Termination Criterion: Condition that determines whether a fixed solution has been found or a stopping condition, for when no solution has been found.

2.2.3 System Architecture

A rule based AI system has five major elements:

- Knowledge Base: Contains knowledge about a specific domain [21]. A knowledge base is made of facts and rules [40].

- Inference Engine: An inference engine seeks information and relationships from the knowledge base and provides recommendations, predictions and answers [41].
- Explanation subsystem: An explanation subsystem is built on top of the "Inference Engine". This component helps analyze the structure of the reasoning performed and explains the decisions made by the system to the user. [21]
- User Interface: The user interface of a rule based expert system consists of Graphical interface that represents the results and explanations (either in Natural Language or Visual Representation) of the reasoning to the 'User' [21].

2.2.4 Inference Techniques

An inference engine parses the rules in the knowledge base and compares it to the facts. By executing rules, a new fact may be found and will be added to the fact memory. Executing many rules for inferring new facts creates "inference chains" [21]. There are 2 main ways an inference engine executes rules [21],[41]:

- Backward Chaining: The process starts with the conclusion/goal and works backwards to find rules and supporting facts that allow the system to achieve its goal.
- Forward Chaining: This process starts with the facts in the knowledge base and works towards achieving new conclusions/goals.

2.2.5 Advantages and Disadvantages

2.2.5.1 Advantages

Expert systems contain the information that has been provided by experts in the a given field. There is traditionally a lower rate of error in the facts and rules and in turn the inferences. Expert systems aim to imitate an expert's decision making, thus reducing the manpower and time required for rudimentary (and in some cases advanced) decision making tasks. These systems also work very well for repetitive tasks and decisions. Rule based expert systems typically work on open world assumption, thus stating that a lack of knowledge implies an 'unknown' state, rather than a false one.

2.2.5.2 Disadvantages

Rule-based AI has a number of limitations, and some of these were major factors in the downfall of expert systems in the early 1990s. As the complexity of the expert systems increase over time, it becomes very difficult to encode the rules for a specific domain. When new facts or rules are added to a large system, it becomes difficult to validate them. Depending on the expert that helped construct the system, there may be a bias in the decisions made. As the domain covered by the expert system gets broader, codifying the rules becomes increasingly difficult. Developing expert systems can take a very long time because of the manpower and time required to build, validate and maintain the knowledge base.

2.3 Situation Calculus

”Situation Calculus is a logical language for representing dynamic changes” [31]. It was introduced by John McCarthy in 1963 [42] and revamped in 1991 by Ray Reiter [43]. The three basic components of situation calculus are *actions, situations and fluents*. In this section we explain the basics of situation calculus including its components, axioms and problems. We will be using a running example of a Robot picking up a box and placing it at an assigned spot.

2.3.1 Different Chains of Thought

McCarthy and Reiter had slightly different views of the Situation Calculus formalization. According to McCarthy (and Hayes), a situation is ”the complete state of the universe at an instance of time” [31],[44]. Reiter on the other hand said that a situation is ”a history, a finite sequence of actions” starting with the initial situation ’ S_0 ’ [45]. While the approaches to the formalization may vary, both McCarthy and Reiter viewed situations ”as first order objects that can be quantified over” [31].

2.3.2 Components

- Actions: ”Actions are what make the dynamic world change from one situation to another” [31]. Actions are performed by an agent.

Examples:

$$pickup(x), move(x, y), putdown(x) \tag{2.1}$$

$$"Poss(a, s) - \text{Action } a \text{ is executable in situation } s" \quad [31]. \quad (2.2)$$

- Situations: According to Ray Reiter, "a situation is a finite sequence of actions (a history)" [45]. According John McCarthy, "a situation is a snapshot of the state of the world at any instance of time" [46].

Examples:

$$S' \leftarrow do(pickup(x), S) \quad (2.3)$$

- Fluents: "Situation-dependent functions used to describe the effects of actions are called fluents" [31]. Fluents can be thought of as properties of the world. Examples:

$$holding(x), onTable(x), on(x, y) \quad (2.4)$$

$$"Holds(p, s) - \text{Fluent } p \text{ holds true in situation } s" \quad (2.5)$$

2.3.3 Axioms

As mentioned earlier, situation calculus is a logical language for reasoning over dynamically changing worlds/domains. Axioms are central to this reasoning, as they enable the system to understand what rules the selected world follows. An application domain can be axiomatized using:

- Action Precondition Axioms: Action precondition axioms are necessary and sufficient conditions that characterize when an action is physically possible [47].

Example [48]:

$$Poss(pickup(x), s) \equiv [(\forall z)\neg holding(z, s)] \wedge nextTo(x, s) \wedge \neg heavy(x) \quad (2.6)$$

- Action effect Axioms: World dynamics are specified by effect axioms. They describe the effect of a given action on the fluents [48].

Example [49]:

$$Poss(drop(r, x), s) \wedge fragile(x, s) \supset broken(x, do(drop(r, x), s)) \quad (2.7)$$

- Foundational Axioms: Foundational or General Axioms are axioms that are true in all situations.

Example:

$$\forall(x, y, s) : on(x, y, s) \wedge \neg(y = Table) \rightarrow \neg clear(y, s) \quad (2.8)$$

2.3.4 The Frame Problem in Situation Calculus

McCarthy and Hayes [44] found that "axiomatizing a dynamic world requires more than action precondition and effect axioms". They deduced that "Frame axioms", which specify the action invariants of the domain, i.e fluents which remain unaffected by a given action [31], are required to reason over situations in a domain. But the number of frame axioms required for a system to reason efficiently is extremely large and sometimes unnecessary. For example, if we are concerned with changing the color of the box the robot picks up, then dropping the box will not change the color of the box, and something as arbitrary as electing a new president will not change the color of the box. Both are frame axioms that would be useful for the machine because they specify which actions won't affect the outcome of the next situation.

2.3.4.1 Reiter's Solution to Frame Problem

Ray Reiter [43] proposed the concept of Successor state axioms as a solution to the 'Frame problem'. Reiter suggests we assume that effect axioms for a specific fluent "describe all the ways an action can change the truth value of the fluent" [48]. "A syntactic transformation is then applied to the effect axioms to obtain successor state axioms" [50].

Example [50]:

$$\begin{aligned} & Poss(a, s) \wedge [(\exists_r)\{a = drop(r, x) \wedge fragile(x, s)\} \vee \\ & (\exists_b)\{a = explode(b) \wedge nextTo(b, x, s)\}] \supset broken(x, do(a, s)) \end{aligned} \quad (2.9)$$

2.3.5 A Common Use Case in Situation Calculus

A typical scenario involving situation calculus starts by providing the system with the 'Initial State' of the world to be modeled. An initial state is a list assertions that can be made about the world before any actions take place. In other words, an initial state is all the assertions and facts about S_0 , the initial situation. Next, a list of actions and events that occur in the modeled world are compiled in chronological order. We then create axioms for

the modeled world, which in turn can be used by the system to infer the next (appropriate) course of action.

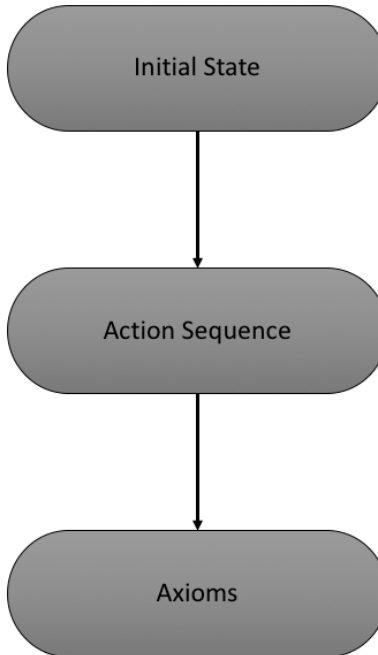


Figure 2.1: Situation calculus workflow. The user starts with an initial state (Situation S_0), and foundational axioms that hold true throughout the narrative. Next, we get a sequence of actions that occur in the narrative, and the fact base comprises mainly of these actions. Lastly, we document axioms/rules based on what we understand about the domain (world presented in the narrative).

2.3.5.1 *GOLOG*

Levesque et. al [48] in their 1997 paper on GOLOG, introduced GOLOG as a logic programming language for dynamic domains. GOLOG’s interpreter ”maintains an explicit representation of the dynamic world being modeled based on user defined axioms” [48]. GOLOG programs can [48]:

- Reason about the state of the world.
- Consider the various possible courses of action before committing to a specific behavior.
- Be written at a higher level of abstraction than is usually possible.

2.4 Deep Learning

”Deep Learning allows computation models to learn features/representations of data with multiple levels of abstractions” [19]. These computational models are composed of multiple processing layers. ”It is part of a broader family of machine learning methods” [51].

2.4.1 History: The Birth, Decline and Resurgence of Neural Networks

Deep Learning can be traced back to a paper by Walter Pitts and Warren McCulloch in 1943. In this paper, the authors discussed how the neurons in the brain work and build a simple model of what can be considered the first neural network [52],[53]. In 1962, Rosenblatt proposed the idea of a training algorithm in artificial neural networks in his book titled, ”Principles of Neurodynamics” [54],[55]. Rosenblatt thought these neural networks, called ”Perceptrons”, were ”potential models of human learning, cognition and memory” [3]. The excitement about the field of ”Artificial Neural Networks” was almost completely silenced by Minsky and Papert’s book ”Perceptrons” [56]. Minsky, whose PhD thesis was focused around one of the earliest connectionist machines [57], lost faith in this approach and began to change the direction of his research [3]. His book detailed multiple scenarios that a perceptron neural network could not handle. This book essentially led to a cutoff in funding for neural network and connectionism related research for almost a decade.

The introduction and popularization of the ”Hopfield Networks” [58] and ”Backpropagation” [59],[60] resuscitated the field of neural networks in the 1980s. As the second AI winter, which was focused around ”expert systems”, started in the late 80s and early 90s, the field of neural networks was slowly gaining traction and interest. With the advent of the 21st century, the amount of data available to train these neural networks increased. The introduction and successful applications of the multilayer perceptrons and support vector machines greatly aided the revival of this branch of AI, which was by now classified as a separate branch called ”machine learning”. By the 1990s and early 2000s, the field of machine learning changed its focus from the original goal of AI, which was learning in the context of intelligent systems, to learning with an emphasis on competent learning algorithms which borrowed from statistical methods and probability theory [61].

With the rise of ”Big Data”, some of the flaws in ”traditional” neural networks began to appear. Traditional Neural Networks, which used the smallest number of layers [62], were prone to problems like ”Overfitting” [63]. On the other hand, the last few years have seen the

”Deep Learning” approach come to the forefront of AI research. Researchers have started applying Deep Neural Networks to solve problems in many areas.

2.4.2 Recurrent Neural Networks

The first successful application of recurrent neural networks was the Hopfield Network [58]. Recurrent Neural Networks (RNNs) are very useful when applied to data with sequential inputs, like natural language text or speech. RNNs process the inputs one unit at a time. They have a ’state vector’ in their hidden layer units, which stores information about the history of the sequence [19]. In normal feed forward neural networks, the history of a unit is represented by the context of $N - 1$ works. In RNNs, the history is represented by neurons with recurrent connections, thus creating an unlimited history [64]. RNNs can form a short term memory, so they handle position invariance better than traditional feed forward networks.

2.4.2.1 Long Short Term Memory

While RNNs are very useful in training natural language and other sequential data, they are known to be difficult to train. Thus, they don’t always show the full potential of recurrent models. While RNNs theoretically have an unlimited history, the training time increases significantly as the training data and (as a result) the history increase in size. Long Short Term Memory (LSTM) networks address this problem. The first LSTM architecture was proposed in 1997 [65]. Each cell of the LSTM network includes an ”input”, ”output” and a ”forget gate”. Each gate is assigned a value between 0 and 1, with 0 meaning ”*no information is stored*” and 1 meaning ”*all information stored*”. The forget gate decides how much of the history is stored during the training process, based on the value it is assigned [65]. In ideal cases, the forget gate throws away the unnecessary information and train only on the relevant and important information.

2.5 Using Sentence Ordering Algorithms to Arrange a Text Narrative in the Form of a Fact Base

The first and optional step of the SAAG workflow is to order the sentences in the story for constructing a fact base. Ordering sentences is an essential task when dealing with natural language data. Learning to order sentences in an understandable manner is an

innately human task, which generally requires some amount of intuition. Earlier attempts at automated sentence order involved the use of naive algorithms like arranging sentences purely by majority ordering or chronological ordering, but these have clearly not delivered the desired results [66]. Thus, researchers are actively pursuing new and robust ways to order sentences that can compare with human ordering.

2.5.1 Why is It Important in a Fact Base

In Situation Calculus, "a situation is defined as a history, a finite sequence of actions" [45]. Thus the order of the actions executed play an important role in the inferences that are drawn by the reasoning engine.

In a broader context, while assembling a fact base, it is important to present the facts about a single subject in an order that can be easily interpreted by the machine. One example of this can be seen when developing a prolog fact base or a knowledge base for an ontology from raw text. The use of pronouns is fairly common when humans write about a set of topics referring to one or more subjects. The knowledge base/fact base on the other hand needs to explicitly refer to subjects performing the actions. One way to achieve would be to apply a sentence ordering algorithm.

2.5.2 State of the Art

Sentence Ordering is an important task in text summarization [67]. Usually performed as part of the text summarization [66]–[69], there has recently been work on sentence ordering as a stand alone task [70].

There are 4 major criteria to consider while creating a sentence ordering algorithm [66]:

- **Chronology:** Order sentences in the chronological order. In multidocument summarization, sentences are often ordered by publication date of the original text document [66]–[69]. Only sorting by publication date can be a naive approach, thus adding another sentence ordering criteria becomes essential. Chronological sentence ordering can also be performed by training a model on the occurrence of keywords that indicate the order of a sentence [70].
- **Topical Closeness:** This method involves the grouping of sentences by their topic. In some multidocument summarization applications, researchers use a naive method

where each document is considered as a separate topic [67]. But this method does not work for certain kinds of text corpora, like Asynchronous conversations in social networking sites [71] and corpora of multiple novels etc. For such methods, topical clustering algorithms are designed based on existing clustering methods like k-means clustering [72] and Latent Dirichlet Allocation [72]. In other use cases, there is a need to perform topical clustering/ topical segmentation by using a machine learning method with sentences encoded in a multidimensional vector at either a word level [73] or at a document level [74].

- **Precedence:** One useful method to improve the order of a sentence is to precedence relation in the existing text corpus. The Precedence criterion measures the substitutability of presuppositional information contained in a defined corpus segment [68], which may have been achieved using topical clustering [67].
- **Succession:** The opposite of precedence, the succession criterion assess the coverage of consequent sentences after the defined corpus segment [68].

Chen et. al [70] uses a neural network approach to ordering sentences. The authors use 3 different neural network architectures and compare the results. The 3 sentences encoder used were:

- **Continuous Bag of Words (CBoW):** Mikolov et. al [75] describes a method that averages the embeddings of words of a sentence. The CBoW model uses continuous distribution representation of the context of the words in the corpus. This is how it differs from other standard bag of words models. Training complexity is [75]:

$$Q = N \times D + D \times \log_2(V) \quad (2.10)$$

- **Convolution Neural Networks (CNN):** Convolutional Neural Networks were built "to capture simple features at a higher resolution and convert them to more complex features at a coarser resolution" [76]. Formally sentences are represented as [70]:

$$\mathbf{cov}_k = \phi(\mathbf{W}_{cov}^T (\oplus_{u=0}^{l_f-1} \mathbf{e}_{k+u}) + \mathbf{b}_{cov}) \quad (2.11)$$

$$\mathbf{e} = \max_k \mathbf{cov}_k \quad (2.12)$$

Where $W_{cov} \in \mathbf{R}^{(d \times l_f) \times d_f}$ and $\mathbf{b}_{cov} \in \mathbf{R}^{d_f}$ are training parameters, and $\phi(\cdot)$ is a tanh function. Also, $k = 1, \dots, n_w - l_f + 1$, and l_f and d_f are hyperparameters that indicate filter length and number feature maps respectively [70].

- Long Short Term Memory (LSTM): For a basic explanation of LSTM network’s architecture refer to section 2.4.2.1. Using LSTM networks helps eliminate the gradient vanishing problem in recurrent neural networks(Explained ahead in section 2.5.3.3).

Using the arXiv abstracts dataset³ to train models, the results show that the LSTM networks outperform the other methods by a good margin.

2.5.3 Method

Sentence ordering for fact-bases in symbolic AI vary slightly from more traditional summarization applications. For example, in situation calculus, a chronological order is an important criterion than in triples for an ontology. To order sentences for the creation of a factbase, we propose using a combination of two approaches. We use short stories and novels as the data for our overall proposed research method. This means that there are multiple asynchronous ”topics” focused on many subjects in the text. A combination of topical clustering algorithms (which will handle grouping similar sentences together based on the subjects, predicates and object being addressed in the sentence), and the neural network approach (which orders sentences within those clusters in the appropriate order based on the occurrences of certain words or characters in the sentence) has been used as the first step of the SAAG workflow.

2.5.3.1 Topical Clustering + Improvements

An important step of a sentence ordering algorithm is grouping the sentences in the corpus by their topic. Topical clustering is a technique that is applied in natural language processing, especially to aid in multidocument summarization tasks [67]. But large text corpora like novels and textbooks typically consist of multiple topics in the same document. There has been research for topical clustering in asynchronous conversations (in blogs and

³https://arxiv.org/help/bulk_data

emails) [71]. The algorithm proposed in Joty et. al overcomes the flaws of traditional topical clustering algorithms. For example, using LDA for topical clustering only takes into account the term frequency, where as using the bag of words(BoW) approach does not take into account multi-party conversations [71].

In this contribution we adapt the existing topical clustering algorithms to first divide the text corpus into various topics. Next these text segments are ordered by chronological order. To create well ordered fact base, one of the important aspects is to arrange the facts about every topic in right order (maybe a combination of various ways like chronological, precedence or succession words), but the other essential aspect is to extract the subjects and objects for any pronouns or indefinites by performing dependency parsing within the topic clusters. While both are possible through known NLP techniques [67], we choose to perform the former using a deep learning approach, since we have found this approach to have a higher versatility at identifying factors that contribute to the order of sentences in a narrative.

2.5.3.2 Neural Network Approach

The second step of our method is to order sentences for the construction of a fact base using deep learning to train a model that can understand keywords that are important in predicting the order of a sentence. This requires the creation of a data corpus with sentences in the "correct" order. For this purpose, we choose a text data corpus which follows a logical order similar seen in fact bases and rules bases. Since scientific literature, instruction manuals and textbooks follow a logical sequence of sentences in a similar manner to the contents of fact and rule bases, they have been chosen as the training data for this experiment. Scientific abstracts in particular are a perfect fit for this experiment, since they are already divided topically, and resemble the results of topical clustering. Hence, we have chosen the "arxiv abstract dataset" as the training data for our models⁴.

Since it has been established that the best model for order prediction is to use recurrent neural networks with long short term memory [70], we used this neural network architecture as the baseline and attempted to improve the accuracy of the order prediction, by tweaking the parameters and altering network architecture. As part of the future work, we plan to

⁴https://arxiv.org/help/bulk_data

use capsule networks [77] to see if the accuracy from the convolutional neural networks can be improved.

Encoding words in the text corpus is the first step of training a sentence order prediction model. The word embedding vector’s dimensions will be decided empirically, after testing out a range of possibilities to identify the optimum word vector size. After the word embeddings have been created, the processed dataset is then used as an input for the neural networks.

2.5.3.3 Recurrent Neural Networks + Long Short Term Memory

Recurrent Neural Networks (RNNs) are known to accept sequential inputs, thus their applications have been extremely successful when applied to natural language data [78]–[85]. A basic explanation on the working of RNN is provided in Section 2.4.2.

Long Short Term Memory (LSTM) networks were created to improve RNNs. One disadvantage of RNNs is the vanishing gradient problem [86], where the gradient propagated back through the network either decays or grows exponentially [87]. LSTM overcomes this flaw and enables the neural network to learn long term dependencies. For information on the LSTM architecture refer to section 2.4.2.1. Since LSTM networks have been shown to be the most successful neural network architecture to learn sentence ordering [70], we use it as the benchmark for our experiment.

Experiments - Word-level: We have created a recurrent neural network with LSTM to predict the order sentences in a narrative. We have explored both word and character-level embeddings for the text data. Our model takes a set of sentences as input and predict the order for those sentences. The word-level LSTM model, has been trained on 1000 abstracts and been tested on 300 abstracts. The embedding dimensions for the sentences have been set to 500, while those for each word are set to 300. There are also 1000 hidden layers in this network. We trained this model through 300,500 and 1000 epoch iterations. We have observed an increase in the accuracy of the model as the number epochs. There has also been an increase in the accuracy of the results based on the number of the abstracts included in the training data. Since our dataset has a total of 41,000 abstracts, we gradually increased the number of abstracts in the training data upto 4800 (until we hit a wall with our available computational resources) to observe at what limit does the result come close to the current state of art, which uses 884,000 abstracts to get the best results.

Experiments - Character-level: We have also explored a recurrent neural network with LSTM which accepts sentence inputs with character-level embeddings. This model trains much slower due to the lower embedding level, and thus does not require larger training dimensions. The overall model works identical to the word level model as far as network architecture is concerned. But we tweaked the parameters to get preliminary results and gauge the running time for the model. We are currently training an LSTM model with character embedding dimensions set to 32 and sentence embedding dimensions set to 64. This network also has a much smaller number of hidden layers at 128. We used this architecture to train a model upto 4500 abstracts from 200 to 1000 epochs.

2.5.4 Sentence Ordering Evaluation

The evaluation metrics used for sentence order prediction are "Recall-Oriented Understudy for Gisting Evaluation"(ROUGE) and "Bilingual evaluation understudy"(BLEU). For this experiment we will compare the sentences generated by the different models using the ROUGE-L, ROUGE-N and BLEU metrics.

- ROUGE-N: "Rouge-N is an n-gram recall between a candidate summary and a reference summary" [88],[89]. Rouge-N while used to compare two summaries, can also be used to compare two text corpora with different sentence ordering. Rouge-N is computed as follows [88]:

$$\frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (2.13)$$

- Rouge-L: Rouge-L is calculated based on the Longest Common Sub-sequence (LCS) [90]. "Longest common sub-sequence takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically" [90].
- BLEU: BLEU (bilingual evaluation understudy) is an algorithm originally developed as "a method of automatic machine translation evaluation that was quick, inexpensive, and language-independent" [91]. A BLEU score is calculated by comparing a candidate translation to a good quality reference translation. When are the scores accross te corpus are averaged, we reach the final BLEU score [91].

2.5.5 Results

Our approach using LSTM neural networks to predict the order of the sentences in a narrative produced promising initial results. But there were a few restrictions in the execution of this experiment which prevented training of models that could beat the state of the art models. Reasons for this include lack of robust computational resources to train models for extended periods of time, and as a by product the time for each training cycle taking exponentially longer to produce results. We also ran into a wall with anything more than 5000 abstracts as a training sample, given the computational resources available at our disposal. With the above bottlenecks the results produced by our training models can be seen in Table 2.1.

Table 2.1: Results from the LSTM sentence ordering model.

Embedding	Training Abstracts	Rouge-1	Rouge-2	Rouge-L	BLEU
Word Level	1000	f:0.29521 p:0.30375 r:0.30393	f:0.18191 p:0.18267 r:0.18313	f:0.26942 p:0.27510 r:0.27546	19.53433
Word Level	2500	f:0.30013 p:0.30910 r:0.30850	f:0.18666 p:0.18787 r:0.18784	f:0.27420 p:0.28044 r:0.27987	19.58148
Char Level	2500	f:0.3150 p:0.3237 r:0.3229	f:0.2061 p:0.2074 r:0.2071	f:0.2903 p:0.2957 r:0.2958	21.4815
Word Level	4800	f:0.428711 p:0.43501 r:0.43582	f:0.33390 p:0.33484 r:0.33527	f:0.40424 p:0.40981 r:0.40964	34.6301
Char Level	4500	f:0.31347 p:0.32195 r:0.32375	f:0.197021 p:0.197767 r:0.198859	f:0.28703 p:0.29260 r:0.29425	20.81765

2.5.6 Conclusion and Future Work

The results were slowly improving with every iteration, as the number of training abstracts were increased. But we were unable to beat the state of the art with our model, but the results produced by the model were enough to order the sentences in the narrative into a fact base. Because of the above reasons, the sentence ordering model, while being an important step in the SAAG workflow (depending on the language used in the knowledge

base and rule base), is not considered a contribution to this thesis. Instead, we use the existing models from a state of the neural network LSTM model introduced in Chen et al. (2016) [70].

Given the ability to fund adequate computational resources and thereby reduce training time, future work for the sentence ordering step includes training the LSTM models with more abstracts and testing out the accuracy and effectiveness of the predictions. Along with this, we plan to explore other types of neural network approaches untested for sentence ordering and see how they compare to the state of the art LSTM model. Among the other approaches, we specifically would like to explore creating Capsule Network models for sentence ordering.

2.5.6.1 Capsule Networks

Convolutional neural networks (CNNs) are among the most successful and commonly used network architectures in deep learning applications, which have performed well with image, video and text data [76],[92]–[99]. But the CNN architecture though has been known to have some flaws. CNN models perform very well when classifying images which are similar to the training data, but this performance worsens when the test dataset contains images with significantly different orientations or transformations than what is seen by the model. The Capsule Network (CapsNet) Architecture overcomes this flaw.

”A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part” [77]. The CapsNet architecture introduced in Sabour et. al consists of a shallow neural network containing two convolutional layers and one fully connected layer [77]. The convolution layer is composed of 256, 9×9 kernels with a stride of 1 and ReLU activation [77]. The second convolutional layer contains ”Primary Capsules”, which are the lowest level of multidimensional entities. This capsule layer is composed of 32 channels of convolutional 8D capsules. The last layer has one 16D capsule per class and these capsules are fully connected to the capsules from the second layer. Using CapsNet has pushed the boundaries on what neural networks can do with image data [100]–[103], but there has not been many applications of CapsNet focused around text data. Since CNNs perform well on text data, we make the assumption that CapsNet may also be able to perform assigned tasks like sentence ordering. Thus as part of the future work, we plan to create a CapsNet Architecture to learn sentence ordering from

a selected text corpus and compare the accuracy of the predictions made to the results from the benchmark.

2.6 Information Extraction

”Information Extraction is an area of natural language processing that deals with finding factual information in free text” [104]. ”Information Extraction is the process of identifying within text instances of specified classes of entities and of predications involving these entities” [105]. Even in a limited domain, Information Extraction is a non-trivial task due to the complexity and ambiguity of natural language.

2.6.1 Types of Information Extraction

Information extraction helps create a structured representation of the free text data. The following are broad tasks included in information extraction [104]:

- ”**Named Entity Recognition** addresses the problem of the identification (detection) and classification of predefined types of named entities, such as organizations, persons, place names, temporal expressions , numerical and currency expressions” [104],[106].
- ”**Co-reference Resolution** requires the identification of multiple (co-referring) mentions of the same entity in the text” [104].
- ”**Relation Extraction** is the task of detecting and classifying predefined relationships between entities identified in text” [104].
- ”**Event Extraction** refers to the task of identifying events in free text and deriving detailed and structured information about them, ideally identifying who did what to whom, when, where, through what methods (instruments), and why” [104].

2.6.2 Event Extraction

Event extraction techniques can usually be classified into 3 types, based on the methodology [107]:

- **Data Event Extraction** solely relies on quantitative methods to discover relations [107].
- ”**Knowledge Event Extraction**: Knowledge-driven text mining is based on patterns that express rules representing expert knowledge” [107].

- **Hybrid Event Extraction:** In most applications, it is difficult to stay within the boundaries of either method [107]. Hybrid event extraction combines both techniques to leverage the advantages of each.

2.6.3 Fact Extraction

A fact usually involves multiple entities and a relation to associate the entities. Wang(2016) [108] describes the process of fact extraction as combination of "Entity" and "Relation Extraction". The algorithm described in his thesis, assembles the fact triples from the extracted information. Additionally, Liu et al. describes "an end-to-end approach for deriving triples from natural language text" [109].

CHAPTER 3

CREATING RULES IN TEXT FORM BASED ON A NARRATIVE USING CROWDSOURCING

3.1 Introduction

Jeff Howe coined the term "Crowdsourcing" in 2006, and defined it as "an idea of outsourcing a task that is traditionally performed by an employee to a large group of people in the form of an open call" [110]. Crowdsourcing was used to complete tasks that were deemed easy for humans but were still difficult to "computerize" (automate) [111].

Yuen et. al 2011 [111] presented a taxonomy of crowdsourcing methods in a survey paper. According to the authors, crowdsourcing work has most often focused on computational techniques such as creating applications, algorithms, datasets, and performance analysis for these or other computation systems [111].

Crowdsourcing methods are used to create datasets either as a standalone research output, or used for various applications or data models. Most dataset creation efforts have been focused around entity tagging or annotations.

Crowdsourcing services like Amazon mechanical turk⁵ have been used for NLP tasks with some success [112],[113]. Zooniverse⁶ [114] has also been gaining visibility as a platform for people-powered research. Using such platforms, scientists have been able to create high quality datasets that drives their research [115]. Crowdsourcing the creation of the training dataset for our SAAG workflow is essential in producing a well defined, logical, and relatively complete dataset as compared to those automatically extracted from text, since they miss the implicit rules in the text.

3.2 Crowdsourcing Dataset Creation

Finding datasets that can be used in an automated or semi-automated axiom generation workflow has been an extremely difficult task. Hence, our ongoing work includes the creation of such a dataset. The training data for future supervised axiom generation models involve the creation of both the rule base and the fact base. The fact base would contain a sequence of actions, events and situations that take place in a defined world at a specified instance in

⁵<https://www.mturk.com/>

⁶<https://www.zooniverse.org/>

time. The rule base needs to contain the rules that define future course of action based on the fulfillment of certain conditions within a specified scope.

Data Collection has been a common task in a crowdsourcing experiment [111]. For the purposes of the proposed experiment, crowdsourcing is an essential method to achieve the completion of a set of 'knowledge base - rule base' corpora. The creation of a previously unavailable dataset will be a contribution to the artificial intelligence domain as a whole. This dataset can be used to train models on domain understanding and to find mappings between entities in the knowledge base to those in the rule base (See Chapter 7 Future Work).

3.3 Background

Researchers need to clearly define certain components of crowdsourcing while documenting the methodology. We consider the methods described by Keating and Furberg (2013) [116] as these components.

- Goal of Research: Clearly articulate the aim of the experiment and the reason crowdsourcing is the optimal solution.
- Target Audience: Different crowdsourcing tasks require different types of participants. It is very important to choose a participant qualified to perform the task assigned to them. For example, classification of hurricane images requires an expert in earth science.
- Engagement Mechanisms: Identifying the target audience in turn aids in providing them with the appropriate motivation for participation.
- Technical Platform: The technical platform is an essential step in the crowdsourcing process and is highly dependent on the motivation for the participant to stay engaged.
- Data Quality: It is important to know whether the data collected as a result of the experiment meets the expectations of the researcher and falls in line with the original research goal.

Rosenstiel (2007) introduced a simple model describing the activation of human behaviour called "Motive-Incentive-Activation-Behaviour" (MIAB) model [117]. Participation

in crowdsourcing activities requires influencing participants to take action, and thus follows the MIAB model.

Motivation and Incentives may be linked strongly to each other depending on the use case and the technical platform in consideration. Motivation, which may be intrinsic or extrinsic, can be defined as "the reason for acting or behaving in a particular way" [118]. Intrinsic motivation refers to acting while fully understanding the value of an activity and performing these actions simply for achieving fulfillment through the completion of the task. Extrinsic motivation on the other hand is the scenario where participants need external incentives, for example direct or indirect monetary compensation or other forms of recognition [119]. Crowdsourcing platforms like amazon mechanical turk use monetary compensation as an incentive. This kind of motivation is completely extrinsic and thus it is easier to motivate participants irrespective of the task assigned to them. For this purpose, it becomes important to set an appropriate amount for the compensation. Other crowdsourcing platforms like 'Zooniverse' provide participants with an opportunity at authorship with their contribution to the experiments. Although extrinsic, this type of task requires the participant to be interested both in the incentive and the motivation behind the experiment itself.

Activation refers to the participants' decision to act. This may be a result of the motivation and incentive set by the researchers. Once the participants have taken action, the next important step is to continue to keep the participant engaged in the activity. An example of this would be to provide the participants with a topic of task that interests them or for there to be a game-mechanics or other user interface components to keep the activity going. Sometimes, services like mechanical turk are treated like jobs that pay per task.

In the context of crowdsourcing, behaviour refers to setting and if possible controlling the desired outcomes of the crowdsourcing task. This part of the MIAB model also involves setting the specific standards for what results are considered "acceptable" and thus decides which participants are rewarded, since one of the issues with crowdsourcing is participants that simply finish their task with very low standards and accuracy and have very little or no influence on the results.

3.3.1 Data

For this experiment, we use a collection of short stories from Project Gutenberg⁷. Project Gutenberg is a library of over 60,000 free eBooks, the vast majority of which are in the public domain in the US. No permission is needed to credit cite or link to Project Gutenberg as the source, even for commercial use. For this experiment, we have used short stories from Project Gutenberg scraped into 2 corpora:

- Gutenberg Dataset: "A corpus of 3,036 digitized books written by 142 authors from Project Gutenberg" [30].
- Kaggle Project Gutenberg Dataset: A collection of 1,002 stories compiled by Kaggle user Shubh Chatterjee⁸.

From a combination of 4,038 stories, we created a subset of short stories based on the character count, since we wanted to keep the reading time for this stories under 20 mins. When choosing stories of 40,000 characters or less, we have selected a total of 668 short stories for the crowdsourcing experiment. 100 of these stories have been randomly chosen to publish on Amazon mechanical turk as part of the crowdsourcing experiment.

3.3.2 Limitations

Crowdsourcing, while known to create a high-quality dataset relatively inexpensively [115], does have a few pitfalls considering the specific goal of our dataset creation task. Creating rules for a GOFAI system requires :

- knowledge about symbolic AI,
- knowledge about the language in which the rule base is written, and
- comprehensive knowledge of the world defined in the fact base.

3.3.2.1 Solutions

In order to overcome these issues, we have decided to use novels and short stories as the domain of choice for this dataset creation task. Novels and short stories typically create fantasy world that do not need any background knowledge for the average crowd-sourced

⁷<https://www.gutenberg.org/>

⁸<https://www.kaggle.com/shubchat/1002-short-stories-from-project-guttenberg>

researcher to understand the world. Thus, it becomes easy for them to write rules that bind the actions and situations that occur in that world. Another benefit of using novels and short stories is that every story would typically contain different rules from each other. For example, in certain stories humans can fly, while in other dragons exist, and in some other animals can talk. Our hypothesis is that such diversity in the rules of various 'worlds' would help generalize the training of the model and thus prevent overfitting by constantly generating rules with an extremely small world as the setting.

Another way to circumvent writing rules for an expert system, which has a steep learning curve, would be to :

- Set some constraints in the form of the rules
- Have the user write the rules in plain text form.

3.4 Method

While many non-experts may not be knowledgeable about writing executable axioms in various languages like prolog, RDF/XML etc. A crowdsourcing participant who knows about the world defined in a novel or short story, can define the rules that govern that world. The easiest way for the participant to document these rules is in natural language form. Thus, as part of this crowdsourcing dataset creation task, our plan is to :

- Give the user a set of short stories.
- Ask the user to write a set of rules for each story. (We will provide the user with detailed examples of rules written for other short stories that they may be used as reference to create these rules.)
- We then process the contents of the short stories into a factbase, using the technique described in the section.
- We also process the rules base by separating out the content of the rules into the antecedent and consequent and later deducing the relationship between the concepts used in the rule.

3.5 Incentive

Mechanical turk works on a monetary incentive system. The amount of the money provided per task, is decided by the researcher and the participant gets compensated through the portal. Mechanical turk charges a commission fee which adds upto 20% of the fee assigned by the researcher. This fee doubles when the number of assignments per participant is above 10^9 .

The crowdsourcing experiment designed as part of this thesis, involves a task with 2 main parts. The first part is to read a short story presented to the participant. After the participant has finished reading the story, the next part of the task is to document the rules for the domain/world described in the story. The detailed explanation for what these rules are and how they should be written will be provided to the participant, with some examples. Since the task is fairly simple and does not require any expertise from the participant, there is no specific age or qualifications required for the participants of this experiment. While simple, this task maybe slightly time consuming, depending on how long it takes the participant to read the short story. For these reasons, we have set an incentive of upto \$0.50 as a reward for each task.

3.5.1 Incentive Adjustment

The requirements for the successful completed task have been decided and documented in this section. The stories selected are fairly simple to read and do not require any expertise on the part of participants. Also many of the stories are fairly well known, since they are in the public domain and have been published a long time ago. If the participant acknowledges that they have already read the story in the past, then they may skip the first part, but they will only receive $1/5^{th}$ of the compensation promised.

The overall compensation has been decided, depending on the length of the story selected randomly for the participant. This is an optimal way of handling the incentive, since the time required to read the stories and the resultant task of writing the rules vary significantly based on the story length.

⁹<https://requester.mturk.com/pricing>

The requirement for the obtaining the full compensation is a threshold of 10000 characters. The exact compensation can be decided based on the following formula:

$$Comp = \frac{N_{char}}{Char_{max}} \cdot Comp_{max} \quad (3.1)$$

Where,

$Comp$ = Calculated compensation,

$Comp_{max}$ = Maximum Compensation decided by the researchers,

N_{char} = Number of characters in the story,

$Char_{max}$ = Maximum characters set as threshold.

3.6 Curation and Evaluation

Datasets collected by crowdsourcing are high in quality. It has been observed that variance in the compensation amount does not significantly affect the quality of the results [115],[120]. The results gathered by this experiment need to be curated and then evaluated for its use in the next step of this thesis, since the main goal of this thesis is to create rule base for a given knowledge base. The rules defined by the participant are in plain text form and will be converted into their logical representations in the next part of this thesis. Thus the evaluation has to focus on the accuracy of the statements in creating a rule base.

Evaluating a set of rules written by humans for a particular topic is qualitative in nature. One potential method is to crowdsource the evaluation of the rules created by another participant. Crowdsourcing evaluation methods is a fairly common method [121]–[124]. Other methods used include [125]:

- Expert Evaluation : A good way to evaluate the output from the experiment is to have a domain expert curate the data. Although it is a method to obtain accurate results, it is hard to scale up this approach, since there has to be a large number of experts to evaluate the results from crowdsourcing experiments.
- Voting/Rating Mechanism : The scalability issue can be resolved by using a voting or rating mechanism for other participants in the crowdsourcing experiment. While

overcoming the scalability issue, this does reduce the reliability of the evaluation, since the participants don't necessarily have expertise in the domain used in the experiment.

- Third party evaluation : Certain third party organizations help evaluate the quality of results. These organizations offer detailed evaluation of the results but may take a long time to produce results and they would cost additional money, which cannot be afforded by many researchers.

We propose the evaluation of results obtained from the crowdsourcing experiment be conducted in two parts. First an expert evaluation of the results has been conducted on the text results of the amazon mechanical turk HITs(Human Intelligence Tasks). This evaluation step is conducted to accept or reject entries submitted by mechanical turk (mturk) workers. An expert evaluation of the rules produced by the mturk workers is based on a few guidelines and examples set by the requester at the beginning of the experiment, along with examples of both good and bad results. A few of the categories examined for this expert evaluation are:

- Rule form: We look for rules broadly in 2 forms. A "simple" rule, where a condition/fluent is stated to be true or false for the given world. And a "complex" rule, where a conditional trigger is presented and the result of how the world is affected by that trigger being fulfilled is also document.
- Grammar and Mistakes: We look for rules with good sentence construction and correct grammar and no spelling mistakes or incomplete sentences.
- Specificity: Another important aspect considered in approving mturks HITs is the usage of concepts/agents, actions and relations in the rule. It is best for the rule to explicitly mention the concepts used in the knowledge base and avoid use of ambiguous pronouns and indefinite words.

We rejected HITs that do not meet the 3 criteria mentioned above. Other than the criteria which apply to most entries, there were other submissions that were immediately rejected because they submitted rules that only contained garbage phrases, or direct excerpts copy-pasted from the story. The results from the first batch published indicated that out of the first 60 HITs, 28 had to be rejected for not meeting these criteria. But with the lessons

learnt from this batch we updated the instructions and have seen a higher amount of success in the submissions since then.

Since the major purpose of the evaluation step is to create rules for a knowledge base, we propose that the evaluation method should be focused on that aspect. The second step of the evaluation for the rules will be conducted after the text results of the crowdsourcing experiment has been converted to formal rules. Detailed explanation of the evaluation method and metrics used will be documented in the Chapter 6.

3.7 Potential Issues

Crowdsourcing works on the principles of:

- dividing a large task into small parts and combining the individual results into achieving a large and high quality result, and,
- the combined opinion of large group of participants is typically better than that of one. It also overcomes the potential biases of a single researcher.

But crowdsourcing is known to sometimes return vast amounts of noise that maybe of little relevance to the task [125],[126]. Quality checks and evaluation methods as described in the previous section each attempt to overcome some parts of the issues of crowdsourcing, but do not solve all of them. Thus mixed method approaches are preferred, and these methods are usually customized to the need of the individual projects and experiments.

3.8 Results and Conclusions

We successfully conducted a crowdsourcing experiment where amazon mechanical turk participants were asked to write rules for a given story. We used 65 randomly selected stories from the databases described in section 3.3.1. 3 participants were asked to write rules for each story. Each participant was asked to write a minimum of 5 rules, and a maximum of 20 rules for each story.

These results were meant to be processed into formal situation calculus rules using NLP based techniques. However, there were many inconsistencies in these results that led to difficulties in processing these sentences into situation calculus rules. We thus had to examine these inconsistencies, elaborate on their causes and how to fix these rules so that

they may be useful in creating formal situation calculus rules. In chapter 4, we explore why the results of our crowdsourcing experiment were inconsistent, what were the inconsistencies in these results, what was our takeaway from running this experiment, and what we did to fix these inconsistencies. The exploration we conducted based on the inconsistencies of the crowdsourcing experiment actually led to some interesting findings which is a contribution to this thesis.

CHAPTER 4

OBSERVATIONS ON AUTOMATED AXIOM GENERATION WHEN CROWDSOURCING IS AN INADEQUATE PROXY FOR HUMAN EXPERT IN THE LOOP

4.1 Background

In this thesis, we create and execute a semi-automated method for axiom generation. The 3 major components of this method are: 1) A sentence ordering model to arrange sentences in the data similar to those seen in a situation calculus knowledge base. 2) Crowdsourcing based method for writing rules in text form. 3) Using NLP and Deep Learning methods to extract formal rules from the results of crowdsourcing.

The original intention of using crowdsourcing for the axiom creation was 2-fold. 1) There exists no "Rule base, Knowledge base" combination dataset to train a completely automated model. 2) In order to create such a dataset, we need human intelligence as part of a method that is not completely manual.

But general crowdsourcing services have participants who do not have expertise in:

1. Symbolic AI
2. Situation calculus (or any language in which axioms are created)
3. The domain defined in commonly used knowledge bases.

We overcome this by:

1. Set some constraints in the form of the rules.
2. Have the user write the rules in plain text form.
3. And most importantly choosing short stories and novels as the domain for axiom generation experiment.

4.2 Observations (Lessons Learnt)

After conducting the experiment to use crowdsourcing to obtain rules for a narrative, a few interesting results came to light. This chapter covers the results of the crowdsourcing

experiment, an evaluation of these results and observations of these results. Based on these results we also make a case for crowdsourcing being an inadequate proxy for human experts, even in cases where domain expertise is not required.

4.2.1 Rules from the Crowdsourcing Experiment are Inconsistent

While developing the workflow for crowdsourcing experiment, potential sources of uncertainty were identified and solutions addressing these sources of uncertainty were documented in Section 3.3.2, 3.3.2.1 and 3.7. But even with precautions in place for different rule outcomes (i.e. contradictory rules), according to the criteria set forth at the workflow design stage was the grammatical and linguistic consistency of the rules written by the participants of crowdsourcing experiment. This assumption however failed to meet the standards required for the larger thesis methodology. The resultant rules show inconsistencies in references to instances, classes, events etc. While accommodations for the traceback of these occurrences were made by inclusion for conference resolution, lemmatization and spell check functions in the work following the results of the crowdsourcing experiment and its NLP based extraction, the inconsistencies in the linguistic and grammatical variances (and many a time errors) making it very hard to generalize and create an automated or even semi-automated solution for this problem. This along the with sentence formatting errors seen in the results of the crowdsourcing experiments, like missing punctuations or incomplete sentences, make NLP based extraction even harder because they result in confusing and sometimes misleading dependency parsing results. For example, "had the author never met Mr. Hammer the event would have never occurred in the first place." results in a dependency parse with 2 objects and no subjects in the LHS of the complex rule. The addition of one ',' fixes this issue. The correct sentence would read as "Had the author never met Mr. Hammer, the event would have never occurred in the first place."

Human experts performing this task would not be required to write the rules in natural language text, and could skip straight to writing rules required in the language of their choice. Thus saving time and effort in the longer run.

4.2.2 Curation Takes a Long Time

An unexpected outcome of this experiment was just how long it took to curate the results. Curators need to at least skim the story before or during the curation and rule

evaluation process. They also need to check if the rules fit certain criteria (described in Section 3.6) and individually approve each of these attempts. Additionally, given how amazon mechanical turk’s workflow, paying out bonus is an additional step that needs to be streamlined. Given the lack of record keeping in the mturk UI, this is a time-consuming process. For further details on length of human evaluation of rules check section 4.3.2.1.

4.2.3 Axiom Boundaries are Difficult to Frame

The frame problem is one of the oldest ones in situation calculus (see Section 2.3.4), and there are solutions available. But given our attempt at (semi-)automation and generating axioms through a non-traditional method, this problem resurfaces, and future work will need to include a way to mitigate this problem.

One of the key explorations in the future work section of this thesis (see Section 7.4) is a deep learning-based axiom generation model. Using this method, the frame axioms would ideally be included in the training data. But since there is no such data available, and we are using crowdsourcing to first create this dataset for the AI community at large to attempt to explore this area. But in such a case solving the frame problem becomes important. And because of the novelty of the method for the creation of this dataset, this yet unknown solution will also be a novel contribution. It may be based off Ray Reiter’s existing solution but needs to be adapted to fit into an automation or crowdsourcing based workflow. The evaluation may need to address this as well as part of the future work, especially in the situation calculus domain.

4.2.4 Rule Extraction Has High Complexity

”Even in a limited domain, Information Extraction is a non-trivial task due to the complexity and ambiguity of natural language” [104] (for more details see. Section 2.6). Rule extraction, a type of IE, applies the principles as entity extraction, but is a more complex endeavour, even compared to applications in triple extraction for knowledge base construction [108],[109]. Rule extraction has additional factors to consider in its identification of the clauses, entities in the clauses, and relationships between the entities in each clause and the overall relationship between clauses. This may be one of the reasons rules extraction has not been attempted many times in the past, and even in the cases where rule extraction has been attempted and been successfully performed, most of the work in this area focuses on

extracting rules from text documents with formal language, that explicitly mention rules in the text [22]–[29]. The concept of finding factual information from **free text** is still a complex problem to solve because of the way humans generally write rules in natural language. When given the option to write rules in an informal language, without the requirement of expertise in the framing of rules (be it for writing axioms for a symbolic AI system or for a medical or legal document). The formal nature of a document and the writing style makes it possible to identify and extract patterns that emerge as a result. But since crowdsourcing participants do not follow such patterns or fixed templates or writing styles, it is hard to create a generalized method for the extraction of these rules.

4.2.5 Considered Choice of Fictional Stories and Its Potential Limitations

A cognizant decision was made to choose fictional stories from Project Gutenberg as the data for our SAAG workflow. This was done because each fictional story would be its own self-contained domain, and there may be limited chance of personal bias slipping into such fictional settings. For example, a non-fiction book like the "Art of War"¹⁰ contains rules stated in the document explicitly and may be used to directly extract rules, but the understanding, but people with prior knowledge and/or expertise in the subject may be much more adept at identifying and writing rules for such books, which does not provide a level playing for the crowdsourcing participants. Additionally, non-fictional books will result in rules that document our current world and thus will create one large rule base that will be difficult to scope down, because we are left with a very large number of rules most of which will not be useful in reasoning for a given story.

However, choosing fictional books introduces its own uncertainties, as seen in the lessons learnt section above (see section 4.2). The complexity of these stories made it harder to evaluate the accuracy of the rules independently and made it difficult to identify the boundaries of the domain. Both the complexity of the rules and the lack of boundaries for the rule base are discussed in detail in sections 4.2.4 and 4.2.3 respectively.

4.3 Human Evaluation of Crowdsourcing Results

Since the rule outputs from mechanical turk are inconsistent and have grammatical and other errors, we need to evaluate and fix these errors in the dataset. Since these errors can not

¹⁰<https://www.gutenberg.org/files/132/132-h/132-h.htm>

be generalized into any patterns to automate the correction of these errors, we developed a manual evaluation method. This evaluation will be focused on the rules produced in natural language text form. The evaluators will not be responsible for the conversion from English (Natural Language) to Situation Calculus.

4.3.1 Evaluation Criteria

In this section we expand on the evaluation criteria briefly mentioned in section 3.6. Our evaluation method observed each rule at 3 levels. If all the criteria in each level are met, the appropriate label will be added and the evaluator will move on to the next level.

4.3.1.1 Level 1: Acceptable Criteria for Rules

Rule form: Evaluators will look for rules broadly in 2 forms. A **"simple"** rule, where a condition/fluent is stated to be true or false for the given world (i.e. A simple rule is usually a statement that holds true or false in the story). For example, "There is no sound in a vacuum." And a **"complex"** rule, where a conditional trigger is presented and the result of how the world is affected by that trigger being fulfilled is also document (i.e. A complex rules are conditional rules, with "if-then" or other conditions, and can be divided in 2 separate clauses). For example, "If the wolf sees pigs, then wolf will chase after the pigs." Evaluators will mark the rule as "simple" or "complex" based on the above criteria.

Specificity: Another important aspect considered in evaluating rules is the usage of concepts/agents, actions and relations in the rule. It is best for the rule to explicitly mention the concepts used seen in the knowledge base and avoid use of ambiguous pronouns and indefinite words.

If there is context for what concept the pronoun is referring to, then the rule is acceptable. For example, "If the wolf sees pigs, then he will chase after the pigs.", "if abby is free, she will feel better."

If there is no context for the concept being referred to with the pronouns, the rule should be marked as rejected. For example, "If he doesn't see the stop sign, he will continue driving.", "They play football every friday."

4.3.1.2 Level 2: Fixing Grammatical Errors and Sentence Construction Issues

Evaluators look for rules with good sentence construction and correct grammar and no spelling mistakes or incomplete sentences. There are 3 types of conclusions for evaluators checking rules for grammar and sentence construction.

- If the rule can easily be understood and needs minor changes including spelling/punctuations or basic sentence construction fixes, then evaluators correct the sentence and mark it as changed.
- If the rule seems difficult to understand or correct, then evaluators mark it as unusable.
- If evaluators find the rule difficult to understand based on the sentence construction but the rule seems useful (i.e more context is needed for evaluators to approve or change the rule), then they mark the rule as "Undecided" in the Fixed Sentence column (indicating an intention to use the rule, but that the evaluators are unclear on how to fix the grammatical issues in the sentence).

4.3.1.3 Level 3: Labeling the Commonsense Rules

One advantage of "Humans in the loop" methods is the ability to generate or identify common sense statements, or even common sense rules. Creating programs that "have common sense" has been a goal of AI researchers for a long time [127]. According to McCarthy (1960) [127], "A program has common sense if it automatically deduces itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows." Researchers have been trying to create models that are able to identify common sense since the 1960s. But automating this process still remains a significant challenge artificial intelligence. Identification of common sense rules during the evaluation method described in this chapter proves that crowdsourcing methods (and by extension methods with humans in the loop) can be used to generate common sense rules, which when converted into situation calculus show that common sense rules for a symbolic AI system can be generated using the semi-automated method proposed in this thesis. In addition to proving the claims made in the this section, these common sense rules can be used, either in its natural language form, or in its situation calculus form, as part of a rule base during the designing other common sense reasoning systems. For achieving the above results, we have added a third level in the evaluation method. This step involves asking evaluators to label the crowdsourcing results as

”CSR” (Common Sense Rule) or ”NCSR” (Not a Common Sense Rule). Since the evaluator does not need to read the story from which these rules are generated, the CSR labeling is done based on whether the rule is understood by the evaluator without having read the story or without being provided any context for the story. If the rule does not follow common sense, or the evaluator needs to actually read the story to verify the validity of the rule, then the rule is marked as ”NCSR”.

4.3.2 Observations

4.3.2.1 Overview

For evaluating the crowdsourcing results, the outputs from amazon mechanical turk were split into 9 batches with 20 rules sets each. Time taken to complete the evaluation ranged between 1 to 3 hours. 7 of the 9 evaluators reported that the entire evaluation took around 1 hour.

The results from the crowdsourcing evaluation are summarized in Figure 4.1. In figure 4.1, we can see the distribution of accepted and rejected rules, and the comparison of those rules with their properties (i.e. Simple/Complex and CommonSense/Non-CommonSense Rules). When these properties are rendered in a Mosaic Plot, we see that a Majority of Rules that were created in the crowdsourcing experiment were simple rules. Also, for the divide between the CommonSense Rules(CSR) and Non-CommonSense Rules (NCSR), we can see that a majority of the Simple Rules do contain common sense statements. The complex rules on the other hand are even divided contain approximately the same amount of CSR and NCSRs. Additionally, we can see that there is a positive correlation between 'Simple and CSR' and 'Complex and NCSR'. Inversely, there is a negative correlation between 'Complex and CSR' and a highly negative correlation between 'Simple and NCSR'.

The accepted rules are summarized in Figure 4.2. In figure 4.2, we can see that most of the simple rules did not need to be fixed during the evaluation process. The complex rules, especially the NCSRs had errors and needed their sentence construction fixed.

4.3.2.2 Individual Rule Observations and Limitations

While the majority of the corrections involve addition of missing articles, spelling mistakes and fixing typos, there are some interesting examples of ”fixed rules” that highlight

the limitations of the crowdsourcing approach and of an objective evaluation method like the one described in this chapter, section 4.3.

A pattern seen in the rules produced during the crowdsourcing are an inconsistency in the concepts being referred to in the rules. For example in a set of rules referring in the story "A gift from earth"¹¹, the residents of the planet of Zur are referred to in a few different ways. The rules "Zurians and Humans are identical physically.", "If the Earthmen come, the people of Zur will talk about them.", "If anyone on Zur acquires metal, then he or she will become rich." each uses a different referring to the same concept. Such inconsistencies become especially difficult to identify or not curated by an expert in the domain (knowledge about the story), but also an expert in the writing rules for AI systems. Even if this problem is not fixed at this stage, bringing consistency to these concepts will need a whole new set of rules or relationships added manually to identify to the system that the 3 concepts are one and the same.

Another limitation of rules written using crowdsourcing is the lack of boundaries for the domain for which rules are being written. The rules input for a story will include important information that cannot be directly obtained by the reading story, but it is improbable to estimate how many rules are needed to efficiently run an expert system. While this is normally the case for even traditionally created expert systems (with expert input rules), in such systems experts are able to prioritize the most important concepts and rules that need to be added to the rule base in order to run the system smoothly. With crowdsourcing, even though one may produce a large quantity of rules by creating multiple HIITs for each story, prioritization of these rules is not possible. Thus there is a chance that researcher may end up with a set of unimportant rules, while missing some key rules.

A repeating comment from evaluators of the crowdsourcing results, were that there were multiple instances where a "rule" as output by the crowdsourcing experiment participant actually contained 2 or sometimes 3 distinct rules. These rules were not conditional in nature, and thus could not be classified as a complex rule. In these cases, one can identify the number of rules an output breaks into, but actually splitting these rules into 2 or 3 sentences would be the same rerunning the crowdsourcing step in the scientific workflow, and thus cannot be fixed during the evaluation step. This is because we need to separate the rule generation to rule evaluation, in order to avoid confirmation bias in the final results of

¹¹<http://orion.tw.rpi.edu/~anirudhprabhu/ShortStoriesForMturk/AGiftfromEarth>

the experiment. Examples of such results include "Mark breaks up his time into mornings, afternoons and evenings.", "If bodily damage happens, the body will repair itself quickly and efficiently if people receive the cancer treatment." etc.

The full set of evaluated rules can be found in the appendix in section D.

4.4 Crowdsourcing in the Context of Expert Replacements

Crowdsourcing has been used to complete tasks that are deemed easy for humans, but are still difficult to "computerize" [111]. Crowdsourcing platforms have been used to create high quality data that drives research [115]. Commonly used crowdsourcing tasks in mechanical turk include Surveys, Image Classification, Image Tagging, Sentiment Analysis, Emotion Detection, Audio transcriptions, Search Relevance etc.

Crowdsourcing methods have boasted some successes over the years [112],[113],[115],[128]–[136]. In figure 4.3, we compare the various crowdsourcing applications based on the complexity of the task assigned to the participant and the likelihood of the results being used as part of an automated workflow. The bottom left quadrant of figure 4.3 is empty as expected, since a simple task which is not used in an automated workflow need not employ crowdsourcing to achieve its results.

As seen in the top left quadrant of figure 4.3, most crowdsourcing experiments that have claimed to be a success have been used to in application environments with fixed needs or where there is an urgency in the requirements. For example, tasks like identification of contents of an image, or classifying images, videos or text require decision making in a fixed domain, i.e. the participant identifies that multimedia content given to them belongs to one of n types pre-decided by the requester. Such tasks that create a bounding box for the participants performing their tasks are more commonly used in crowdsourcing and rarely require expertise or formal training in the experiment being performed, or context provided for the larger use case driving the tasks assigned to the participants of the crowdsourcing study. Thus an extrinsic motivation like monetary compensation provided for each completed task is usually enough for these studies. The success of crowdsourcing platforms like Mechanical Turk and others that follow a similar model exemplify the frequency of such studies conducted [115]. These studies are usually part of workflow that involve using these result to be machine read and applied to descriptive, predictive or prescriptive models.

When participants are asked to give their opinions on various topics, like those seen in free text surveys (without fixed options), add complexity to the situation by asking the participants to perform tasks beyond common identification, classification or transcription. In these cases, the participant is asked to provide their opinion on a topic presented to them, and this exhibit their interest, understanding and articulation skills on a specific topic/question. The variance in answer length, specificity and eloquence in these experiments is testament to the complexity of such tasks. Even if the "quality" of the answers provided in such surveys and crowdsourcing experiments does not affect the success of these experiments, because the goal of these experiments is to gain opinions and answers from a wider demographic of participants to somehow explain the views of a population sample about a specific topic or to make generalized policy decisions. In such experiments too, expert opinions are not needed during crowdsourcing stage of the workflow, because the role of the expert is examine and interpret the results gained from the crowdsourcing study. Such experiments can be seen in the bottom right quadrant of figure 4.3. In addition to such tasks, the bottom right quadrant of figure 4.3 also shows tasks like semantic segmentation or image summarization, where the results from crowdsourcing are used as the final result of training model and not as part of an automated workflow.

But there are experiments where crowdsourcing tasks require the participants to exhibit their interest, understanding and articulation, and crowdsourcing study is just a part of a larger computational workflow (as seen in the top right quadrant of figure 4.3). The crowdsourcing experiment performed in this thesis fits this category perfectly. Such experiments which ask participants to perform complex tasks where results affect the outcome of workflow use crowdsourcing to replace expert participation in the experiment. Replacing experts with crowdsourcing can not be done for reasons, such as:

- Experts require a larger monetary compensation for the task compared to general crowdsourcing participants.
- There aren't enough experts to perform the task for the study.
- With only a few experts, there is a chance that some of the expert's views and biases may slip into the system. because of the previous 2 reasons, it is also difficult to aggregate over a large number of rules from a number of experts.

4.4.1 Boundaries of Crowdsourcing Tasks and Their Success

Based on our observations (see section 4.2) from the crowdsourcing experiment detailed in Chapter 3, we can see that applying to crowdsourcing as a replacement to experts-in-the-loop has significant issues, and thus does not produce high quality results. Our assertion based on these findings is that crowdsourcing is an inadequate proxy for experts in the loop, and should be avoided in tasks where participation is considered as expert replacements. This is because the problems in the quality of the data collected from such experiments requires much more work to evaluate, clean and process the data for future tasks in the scientific workflow.

Thus, the success of a crowdsourcing experiment depends on how the task is organized, and the scope of that task in the larger the scientific workflow. Research designing crowdsourcing studies should be cognizant of these boundaries, and how they affect the success of the experiment.

4.5 The "Humans in the Loop" Discussion and "Experts in the Loop" Distinction

Most AI researchers, especially those using machine learning methods, focus on automating predictive models based on the training data available [137]. Automated approaches benefit from and usually require large training datasets. However in many cases there are either very small datasets or largely incomplete (sparse) or undocumented datasets available, and in some rare cases (like this thesis) there are no datasets available for certain explorations. In such cases, building an automated workflow becomes insufficient (and sometimes impossible) to gain scientific insights from the data. Here scientific workflows can benefit from human input to reduce the difficulties faced by lack of usable data for machine learning or other artificial intelligence applications in the fields of Natural Language Processing, Pattern recognition etc where human input can help shape the training data either through labeling or tagging. Workflows that require benefit from human interaction and input are called "Humans in the loop" (HITL) workflow. In literature, HITL has often been conflated to include domain expertise, or expertise in the field of the data being trained, and input or opinions from laypeople who are asked to participate in studies, surveys etc [137]–[139].

While the concept of "Experts in the loop" (EITL), is gradually being used and recognized in the recent years [140],[141], especially in the field of Health Informatics and Medicine [142]–[144]. Researchers have not delved into the various uses, advantages and limitations of experts in the loop. Infact, most of the literature reviews of HITL approaches, either imply that experts can include domain experts [137], or mention it as fitting under multiple sub-categories of the HITL framework [139]. There needs to be better documentation of the experiments and workflows using EITL, so as to understand under what categories, EITL would be beneficial to the scientific exploration over the broader HITL, or a fully automated workflow. The experiments conducted on crowdsourcing and our examination of its results have shown that there are definitely boundaries of the HITL, and EITL would resolve the issues faced during the crowdsourcing experiment. Defining the entire EITL framework, and documenting the boundaries of the approach, including advantages, limitations and decision making criteria is currently out of scope for this thesis, the future work that has emerged from this experiment is the use this experiment as a starting point to compare HITL and EITL in various applications and various domains.

While fleshing out the role of experts in scientific workflows, it is important to also define (or at least articulate) what an "expert" is, in the context of the research area and the level of automation in the workflow. Weinstein (1993) argues that "there are two kind of experts: those whose expertise is a function of what they know (epistemic expertise) or what they do (performative expertise)" [145]. In the design of modern scientific workflows, the requirements for the type of expertise may include either type identified by Weinstein or a combination of the 2 types. Hence, we propose considering expertise as a spectrum ranging from amateur enthusiast to leading expert. When experts are being chosen during the creation of workflows in an interdisciplinary settings, "the desired outcome of the workflow" and "the time and resources available to train participants in the loop" are key factors that go into deciding the need level of expertise of the participants in a scientific workflow. For example, in the SAAG workflow proposed in this thesis, if the participants may have additional training in NLP or an understanding about the context of the axiom generation workflow, then we may be able to reduce the inconsistencies exhibited in the crowdsourcing results due to the informal language used by the participants in their rules. But that would require significant amount of time and resources invested to train the participants on the above mentioned topics. Thus, it would be a better choice to have experts who already

have formal training in the NLP and knowledge about symbolic AI to participate in the axiom generation workflows, either to write rules or to curate the written rules for a given knowledge base.

4.5.1 Citizen Science Explorations and Their Boundaries

According to Hossain and Kauranen (2015), crowdsourcing applications can be grouped in the following areas [146]:

- Idea Generation : Participants are encouraged "to submit new ideas and the best ideas are selected" [146]. Sometimes "crowds can outperform professionals in many levels of new product ideation" [146],[147].
- Microtasking : Microtasking is a "system in which users are assigned to complete small tasks for monetary or non-monetary rewards [148].
- Open Source Software : In Open Source Software development, continual active assessment of inputs by the community increases software quality" [146].
- Public Participation : "Public participation via crowdsourcing can engage a wide range of people and it can facilitate an open dialogue between citizens and decision-makers" [146],[149],[150].
- Citizen Science : "Citizen science is a form of collaborative research in which the participation of crowds is utilized in solving real-world problems" [146],[151].
- Citizen Journalism : "Citizen journalism is an alternative media, which is turning to the crowds for Internet journalism" [146],[152].
- Wikies : "Wikies facilitate online work in collaborative environments using a global network of volunteers" [146].

Crowdsourcing areas like Citizen Science and Microtasking are good venues for this comparative study of HITL and EITL, because they provide many applications where both experts and non-experts work together as part of the same scientific workflow, or experts and non-experts are used interchangeably for the similar by different researchers. For example, the Carbon Mineral Challenge conducted by the Deep Carbon Observatory (DCO) [153]

used a workflow where citizen scientists would discover minerals and reach out to DCO researcher would direct them to labs where experts would identify the mineral species, document whether a new mineral had been found. And image classification, for example, can be performed by both experts [154] and non-experts (crowdsourcing) [155] based on the content of the images, the urgency of the need, amount of data to be classified and any additional context of the scientific exploration. Similar comparisons, based on the same factors mentioned above, can be made in NLP applications like text tagging, sentiment analysis, document classifications etc.

4.6 Conclusion

The results produced by the crowdsourcing experiment conducted in this thesis produced inconsistent results which could not be used to a workflow where these results need to be processed automatically as part of larger computational workflow. Based on the observations of the results, their flaws, and the lessons learned from the additional evaluation that had to be performed for these results and the fact that the NLP processing step could be entirely avoided if an expert in situation calculus were to axioms for the stories used as the knowledge base for this thesis, we assert that crowdsourcing is an inadequate proxy for experts in the loop, and should be avoided in tasks where participation is considered as expert replacements.

Future work for this exploration includes comparing the EITL workflows to HITL workflows and creating a complete framework for "Expert in the loop" in artificial intelligence, including documenting the boundaries of the approach, including advantages, limitations and decision making criteria.

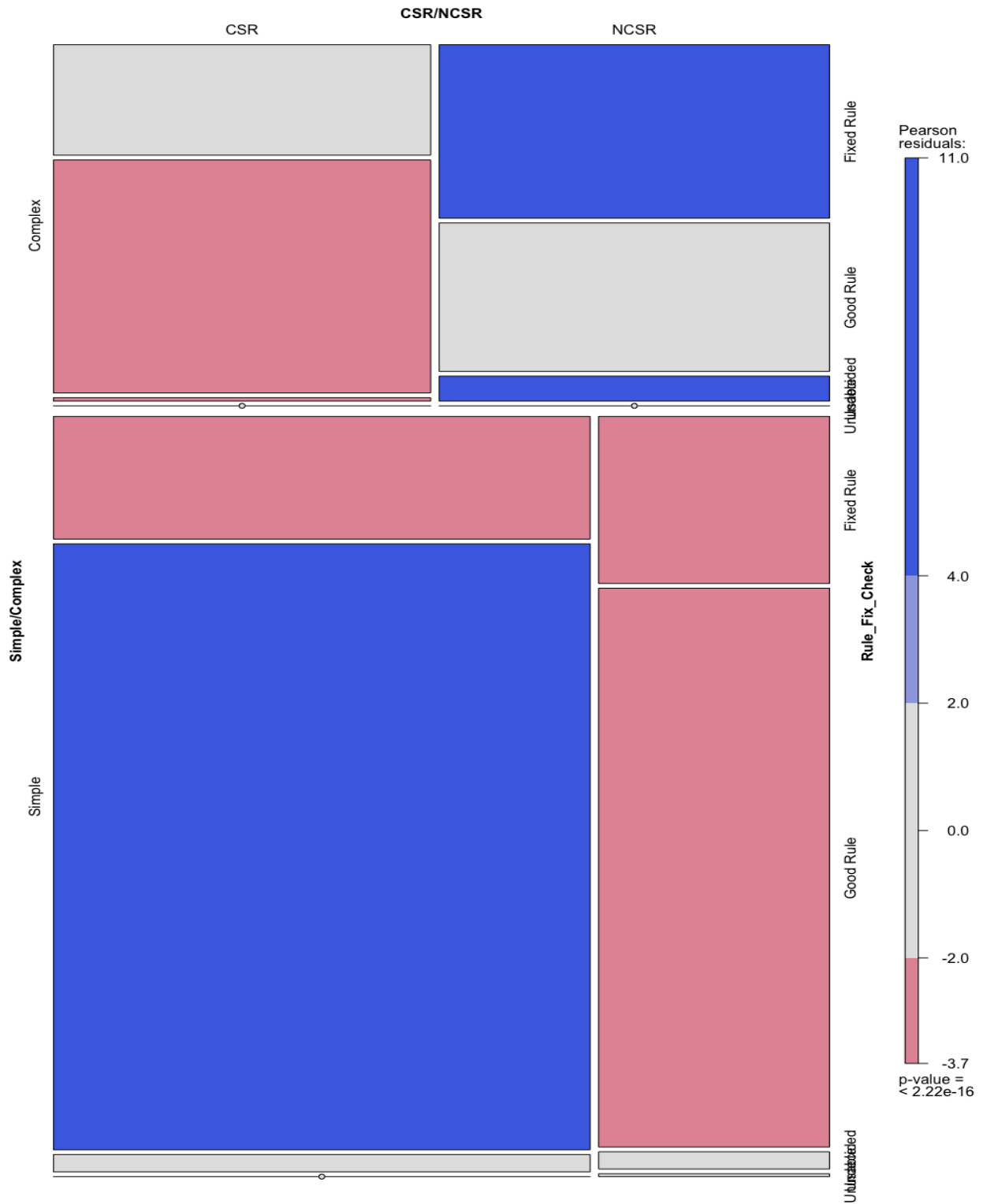


Figure 4.2: Mosaic plot representing the properties of the rules that have undergone the evaluation process as described in Section 4.3.

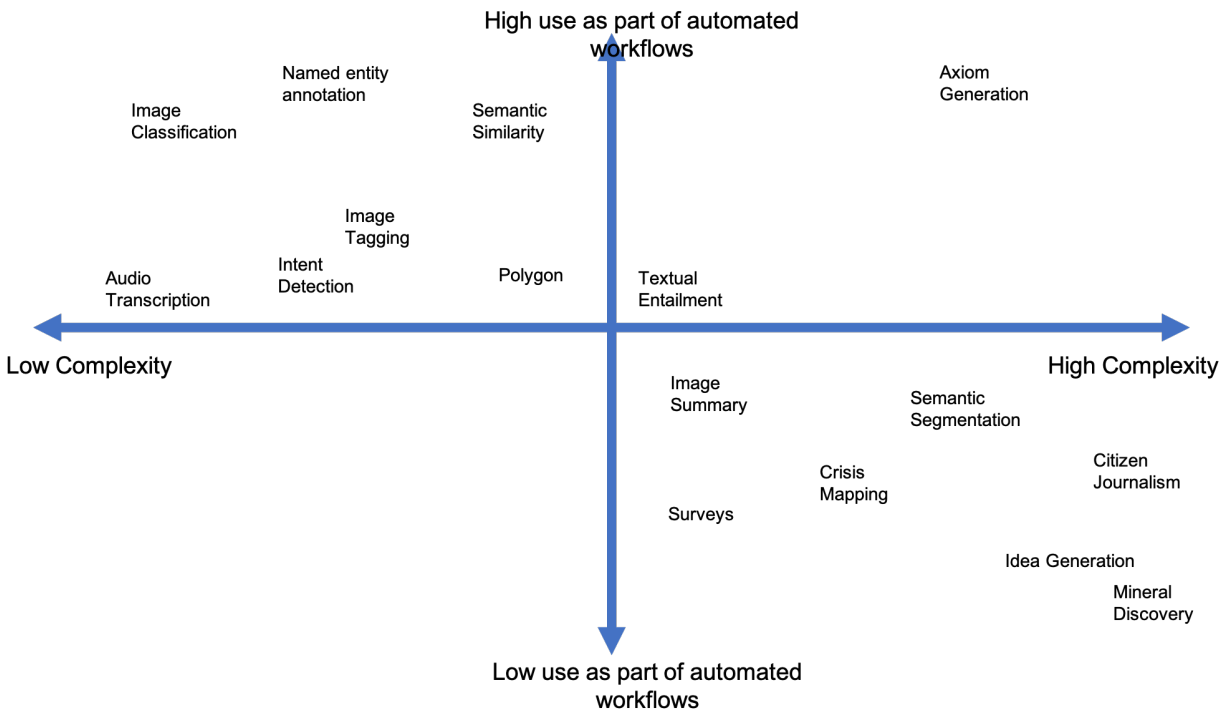


Figure 4.3: Comparison of the crowdsourcing tasks. The X-axis shows complexity and the Y-axis shows the likelihood of the results being used in an automated workflow.

CHAPTER 5

CONVERTING TEXT RESULTS OF THE CROWDSOURCING EXPERIMENT INTO FORMAL RULES AXIOMS

5.1 Background

Text documents contain a plethora of information, enough to create valuable structured datasets. Extensive research has been conducted on triple extraction and knowledge base construction [108],[156]–[160]. Extracting rules from text on the other hand is a more ambitious effort. Rule extraction from text data has been explored since the 1990s [27]. Delanoy et. al (1993) used a machine translation based approach that the authors called MaLTe(Machine Learning from Text) for converting text to rules [27]. Dragoni et. al (2016) combined the linguistic information in WordNet with logic-based dependency extraction [22]. Wyner and Peters (2011) explored a rule based approach to extracting rules from documents [24]. One common trait observed in most of the research projects tackling rule extraction is that they use datasets that explicitly mention rules as part of their text.

While these methods are known to be accurate and somewhat complete in the rules they extract, a lot of information embedded in the data is lost when considering certain 'implicit' rules. Also, most text datasets do not have sentences that describe and contain rules explicitly. Narratives constructed by humans are full of rules that need to be inferred based on context and subtext. The research method defined for rule extraction defined in this thesis achieves the above mentioned goal.

5.2 Algorithm/Methodology

In this thesis, we attempt to convert rules presented to us in an unstructured text format into structured rules in Situation Calculus. Two commonly used approaches to achieve this goal are machine learning and information extraction (NLP) based methods. In this thesis, we have created and executed a workflow based on known NLP techniques for converting raw text into formal situation calculus rules.

5.2.1 NLP Based Methods

NLP based methods parse sentences using language parsers like 'StanfordCoreNLP', 'spaCy' etc. and examine the grammatical and linguistic features of the data to extract concepts and properties to form rules. Rule bases, like fact bases consist of key concepts and relationships between these concepts. Our approach tries to use known NLP techniques like POS (Part of speech) tagging [161], dependency parsing [162], and co-reference resolution [163] to annotate the rules that are available in text form to identify the concepts and relationships that exist in the unstructured text we have obtained from the crowdsourcing experiment. We have used the 'spaCy' pipeline for both these tasks. 'spaCy' is one of the leading open-source NLP parser written in Python [164].

Since the text data has been obtained from a crowdsourcing experiment, there are bound to be sentences with different structures, potentially with some mistakes. Our methodology first parses each sentence and identifies whether the sentence really describes a rule or can be converted into one. If not, we discard the sentence from the rule base. In this section we have created a methodology that accepts text as input and converts them to rules for situation calculus.

5.2.1.1 *How to Handle Rules*

Rules need to have subjects and objects that the rule addresses. Such rules will have two verbs and must have at least one subject-object pair. Using the number of verbs that occur in the rules, we can identify and broadly categorize the rules for our algorithm. As seen in Figure 5.2, there are 2 types of rules we see in the sentences :

- Rules with an antecedent and consequent. While it is not officially a type of rule, we haven chosen to call these complex axioms.
- Rules which contain only one statement which highlight a condition met in the narrative, but which aren't explicitly stated in the text. While it is not officially a type of rule, we have chosen to call these simple axioms.

In addition to identifying the type of rule using the POS tagging method, we provided the same criteria to the humans evaluators in the crowdsourcing evaluation step for them to classify the rules as simple or complex.

5.2.1.2 *Complex Axioms*

In order to successfully convert text to an axiom, we have developed the method shown in Figure 5.3. We first check for any pronouns used in the text, since it does not help identify the concept being addressed by the rule. The sentences with pronouns are replaced with the noun or proper noun using the Co-reference Resolver. If there is no reference found, then it is impossible to link the subject and object to a concept in the knowledge base and thus we must discard the sentence. When the required references are found, we can move on to matching the concepts or entities stated in the rule to those that exist in the knowledge base and discard the sentences that do not contain any concepts from the knowledge base.

For the sentences that are remaining, we separate out the antecedent and consequent. This split is done using clause segmentation, specifically using the `claucy` python package [165],[166]. The antecedent forms the right hand side (RHS) of the rule, while consequent forms the left hand side (LHS) of the rule. These 2 statements are conjugated by a ' \equiv ' or a ' \supset ' sign in the formal representation of situation calculus, but our text to rules conversation makes rules that can be run using `golog` (a prolog implementation of situation calculus). Thus the antecedent and the consequent are separated by ' $:$ ' or ' $-$ '. The `claucy` package is a python implementation of `clausIE` [165], built to be integrated into the '`spaCy`' pipeline, which we use as our comprehensive natural language processing library of choice [164]. We have also used `spacy` for dependency parsing, and to create a pipeline for co-reference resolution package. We have used '`neuralcoref`' python package¹² to perform co-reference resolution for both Complex and Simple Axioms.

5.2.1.3 *Simple Axioms*

Simple Axioms usually are single statements that hold true for the narrative, but are not explicitly mentioned in the narrative. These sentences have one verb and may contain up to one subject-object pair. Our methodology for handling such rules, as shown in Figure 5.4, mainly extracts triples using the '`clausIE`' prepositions as per the '`claucy`' python package implementation [165]. Similar to the complex axioms, we also use '`spacy`', '`neuralcoref`' and '`claucy`' packages to implement the text to rules conversion for simple axioms.

¹²<https://github.com/huggingface/neuralcoref>

5.3 Results

We applied the text to rules extraction method to all the rules accepted during the crowdsourcing evaluation. The crowdsourcing evaluation eliminated some of the issues caused by the inconsistencies in the rules output by the crowdsourcing participants. We have thus successfully converted natural language text to formal situation calculus rules in Golog. Some examples of rules extracted are:

- looked(children, at the sky) :- see(children, a bright star).
- use(Citizens, elevators).
- are(you, a Earthwoman) :- attract(you, Zurian men).
- are(you, an Earthman) :- come(you, from an overpopulated planet).
- eat(you, your own livestock) :- are breaking(you, regulations).
- are connected(those bright islands, by strings of light) :- be(it, an unusual view).

All the rules generated using our text to rules workflow can be found in appendix E. Of the 1233 rules that were accepted after the crowdsourcing evaluation, 99 were discarded because our method was neither able to detect clauses, nor extract triples from the rule. That still meant that 1134 formal situation calculus rules were extracted using our text to rule conversion method. Examining the rules created, we found that a majority of the rules did not contain any syntactical errors and could be read into a golog rule base. But syntactic correctness is just one aspect of validating the rules in a rule base. On closer examination, there were issues with the results produced using our text to rules conversion method. The next sections explore these issues, the causes for these issues, and planned future work to resolve these issues. A more detailed examination of these rules will also be performed in chapter 6.

5.4 Observations and Limitations

Our contribution of converting text to situation calculus focuses on using existing NLP methods and python packages to successfully complete this conversion. But the results produced in this exploration have shown the limitations of our approach and thus the limitations

of the NLP methods and commonly used packages used in text processing tasks and NLP explorations. Broadly categorized, the issues in the extraction fall into 3 categories.

The first limitation we found in the rules is that claucy's triple extraction was not able to extract negative words as part of the triple extraction (especially the predicates). For example, the text rule "Citizens must not use elevators." was extracted as "use(Citizens, elevators).", but the rule must be "not(use(Citizens, elevators))." Another example of the same issue can be found in the text rule, "Love never dies.", which is extracted as "dies(Love)." and must instead be written as "not(dies(Love)).".

The second limitation is the lack of flexibility in the clause segmentation task for dividing the rule into antecedent and consequent. If-then rules perform well for clause segmentation tasks, but if the rule is phrased differently (as is usually the case), the clause segmentation fails to correctly identify the subjects and objects of the rule. For example, the text rule "People who are poor cannot afford meat very often." is extracted as "are(who, poor) :- afford(People, meat)". If a human was performing this extraction, they can easily identify that this rule can also be read as "If people are poor, they cannot afford meat.", which in turn will correctly form the rule, "are(People, poor) :- not(afford(People, meat)).", by accurately identifying the subjects and the negatives in the sentence. Another example of this can be seen in the rule "You must wear a suit to go outside.", which has been extracted as "wear(You, a suit)". If a human was performing this extraction, they can identify that this rule can also be read as "If you want to go outside, you must wear a suit.", which in turn can be formalised as "go(outside):-wear(person, suit)."

The third limitation was found in the execution of the co-reference resolution. There are some errors in the entity being referenced in rules like, "If the character and her Pa and Ma see someone else, the character and her Pa and Ma her see someone else will get scared.", where the object being referenced was incorrectly identified. Another commonly seen mistake is the the entity being replaced does not take into account the pronoun being replaced and its context. This is especially seen when "her/his" is the pronoun used in the sentence. For example, rules like "If the pain cycle is the worst, Miryam bites down on Miryam under lip.", "If Miryam bites down on Miryam under lip, the flesh shows as white as Miryam teeth.", "The Madame gets all The Madame can for The Madame money." and "Davie always gets relieved of consequences for Davie actions.", we can see that the entity being referenced is correctly identified, but the context of their replacement has not been

identified. So when we use the 'neuralcoref' implementation from the 'spaCy' pipeline, we can see that there are mistakes in the replacements. In the above rules, the entity reference replacement should be written as "Miryam's", "Madame's" and "Davie's" respectively. This is an important step to be performed, because without taking into account the context of the replacement, further processing becomes difficult. In all of the rules mentioned above, identifying the subjects of consequent of the rule shows errors, this can be traced back to the limitations of the co-reference resolution method and its built-in entity replacement.

The last limitation is the lack of generalization in rules, where pronouns clearly imply a generalization. In this case, the pronoun needs to be ignored or needs to be replaced with a super class of the subject implied by the pronoun. For example a rule like, "if someone gets no air or heat, they freeze to die." accurately describes the conditional trigger and the effect of that action taking place. But using existing NLP methods to extract a rule from the given sentence does not work, because the above rule fails at the co-reference stage, because there is an implied generalization made here with the words 'someone' and 'them', both references a super class of the concept (like 'person') that is not mentioned in the rule, only in the story.

5.5 Future Work

After examining the results from the text to rules conversion, we noticed some limitations of our method and identified the causes of these limitations. In summary, limitations of our developed method trace back to errors caused by the inadequacies of existing implementations of the known NLP methods when it comes automated axiom generation tasks. Future work for this contribution includes overcoming the limitations of the NLP based method and execution of a deep learning method based on machine translation algorithms. An overarching theme of these limitations shows the difference between humans performing complex tasks and algorithms designed to do the same thing. Algorithms and computational workflows rely on general patterns to automate tasks or solve problems, but in cases where such generalized patterns cannot be identified or the outliers are as numerous as the valid data points, humans performing the task may be the efficient way to solve the problem.

A new exploration to overcome this limitation will be to build a dataset for the text to rules conversion and training a deep neural network model to perform the text to rules conversion. An important point to consider with this exploration is to have an expert

generated text to rule conversion dataset in the programming logic language of choice. With enough variance in the type of rules converted from sentences, the neural network should be able identify the key components of the sentences and combine them to create formal rules. For our future work, we plan to create a sequence to sequence model based on machine translation applications to convert natural language text to formal rules.

5.5.1 Machine Translation

Machine learning, and deep learning methods by extension, use the principles of the machine translation to convert text into a structured format like rules. The premise here is "rules have a defined linguistic structure and the transition of unstructured text to the structured rules can be considered a translation of the unstructured text." Machine translation based approaches may not need to rely on pre-existing knowledge depending on whether the method chosen is supervised or unsupervised. In unsupervised methods, we need to rely on the knowledge extracted from within the text data. In such cases, we use monolingual corpora from 2 different languages and map them to the same latent space and the model learns to reconstruct both languages from this feature space [167].

Most commonly used methods seen though are supervised machine learning methods, where the user provides the the model with parallel data for 2 languages and the model maps the translation from one word onto the other, there by learning to reconstruct the second language [168]–[171]. Our approach for machine translation will follow a supervised approach, with the input being unstructured text and the output is the rule generated from the text.

5.5.1.1 Deep Learning

There are a some considerations to create a neural network to achieve the above task. First and foremost, we need to understand how to represent the text data from the training corpus. Most text based deep learning models train at a word-level [79],[80],[172], but there are certain tasks where character-level [173]–[175], phrase-level [78],[176],[177], or sentence-level [96],[178] representations are essential.

For our experiments, we use both word and character-level embedding, since there is a need for learning the structure of a rule with the positions of their concepts and relationships. Depending on the availability of time, we would also like to explore a knowledge graph

embedding [109],[179] to pick out the concepts in the knowledge base and use those in the rules. This may be a harder problem to solve in situation calculus, especially for our use case. Since there exists no comprehensive knowledge graph like dataset in the domain chosen for this thesis.

Given the success of RNN-LSTM and Seq2Seq models for machine translation tasks [78],[109],[170],[171],[180],[181], we decided to create a text to rules translation with this combination. More details of the implementation will be added after the experiment has been completed.

5.5.1.2 *Evaluation*

BLEU (BiLingual Evaluation Understudy) is a commonly used metric to evaluate the success of a machine translation task [78],[167],[169],[171]. Since BLEU does not fully capture the quality of a translation [181], we explore other metrics to evaluate the results of the text to rules Model. These quantitative metrics combined with the qualitative evaluation metrics described in Chapter 6 will provide an understanding of the quality of the success of the experiment.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for machine translation evaluation that is based on a generalized concept of unigram machine translation and human produced reference translations [182]. It was created to improve upon BLEU. While BLEU seeks correlation at the corpus level, METEOR highlights good correlation at the sentence or segment level. METEOR creates an alignment between 2 strings (the machine translation and the human translation). Banerjee and Lavie (2005) define a "mapping between the unigrams such that every unigram in each string maps to at least one unigram in the other and to no unigrams in the same string" [182].

The Unigram precision (P) and recall (R) for METEOR are computed same as for BLEU [183]:

$$P = \frac{m}{w_t} \tag{5.1}$$

$$R = \frac{m}{w_r} \tag{5.2}$$

Where m is the number of unigrams in the candidate translations that are found in the reference translation, and w_t and w_r are the number of unigrams found in the candidate and reference translations respectively [183].

F_{mean} is computed by combining the precision and recall via "a harmonic mean, with recall weighted 9 times more than precision" [182],[183].

$$F_{mean} = \frac{10PR}{R + 9P} \quad (5.3)$$

In order to take longer segment matches into consideration (more than just unigram matches), METEOR calculates a penalty for an alignment as [182]:

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams_matched} \right)^3 \quad (5.4)$$

Finally, the METEOR score is calculated as [182]:

$$Score = F_{mean} * (1 - Penalty) \quad (5.5)$$

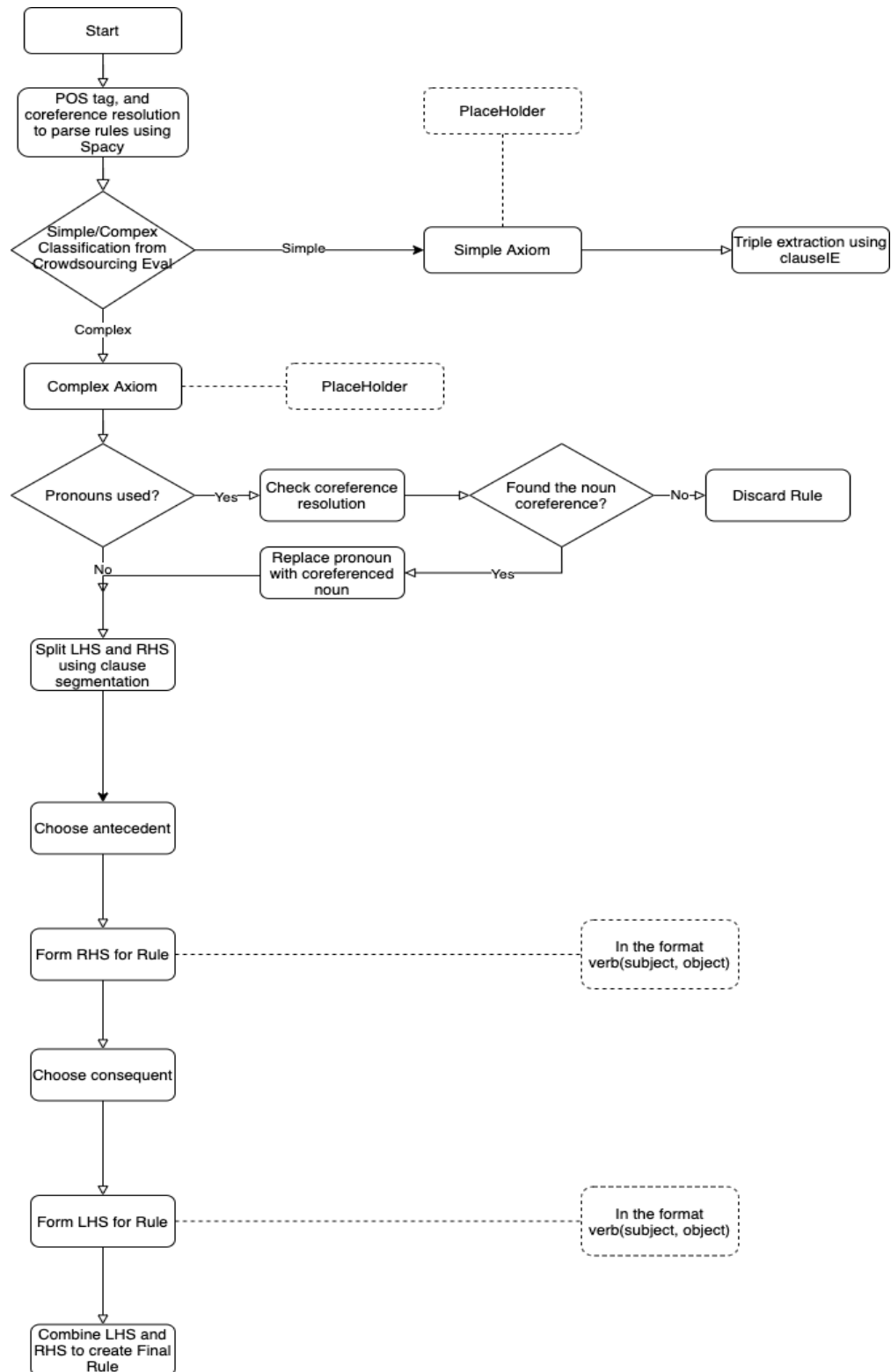


Figure 5.1: Our proposed method for converting results of the crowdsourcing experiment (rules in text form) into formal rules. In the figure, the solid lines and shapes represent the main workflow, and the dotted lines represents comments or notes needed at the corresponding step for easily understanding the method.

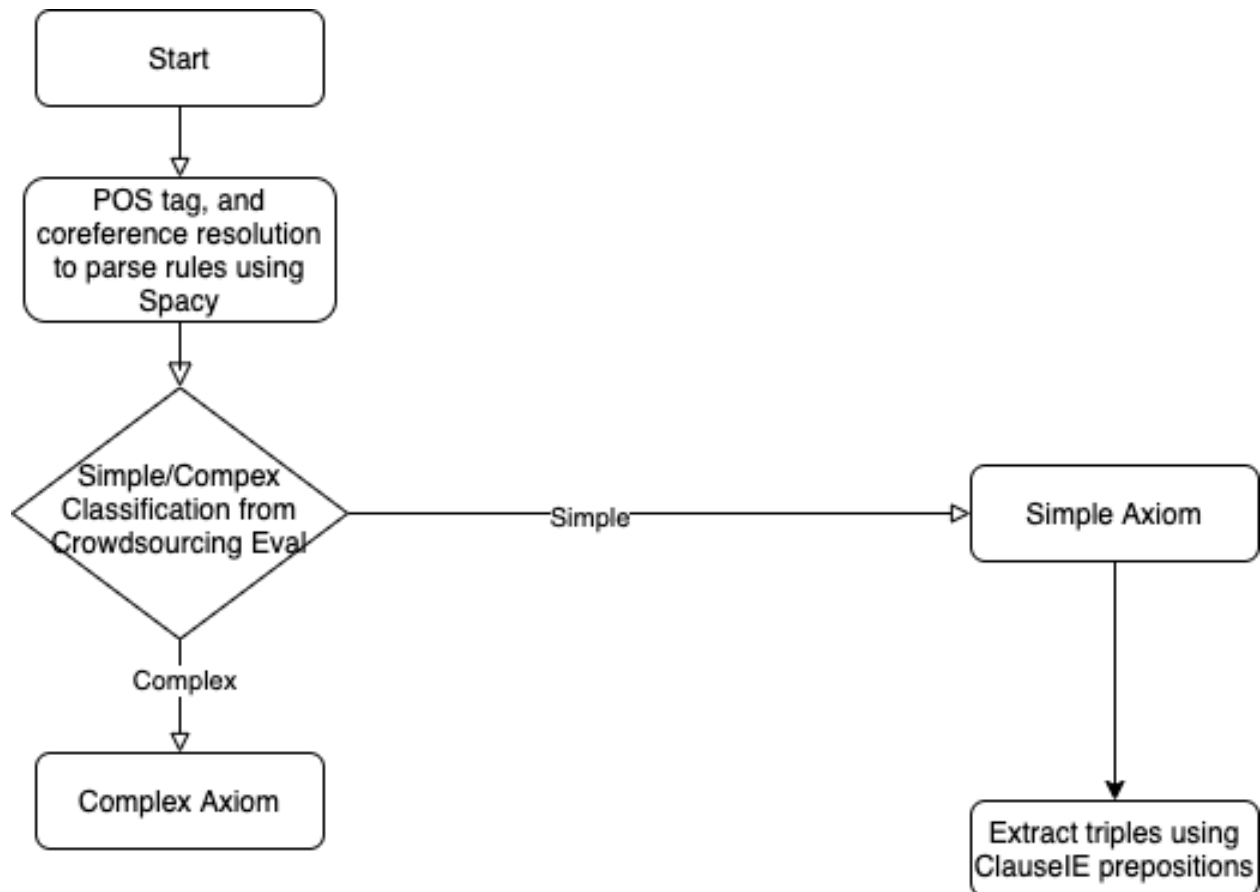


Figure 5.2: We identify the "type" of axiom, based on the results of the crowdsourcing evaluation. The evaluators look for the presence of an antecedent and consequent. A rule with multiple subject object pairs, or a antecedent and consequent is classified as a "complex axiom" by our algorithm. If it meets neither of the 2 criteria, the axiom is considered "simple".

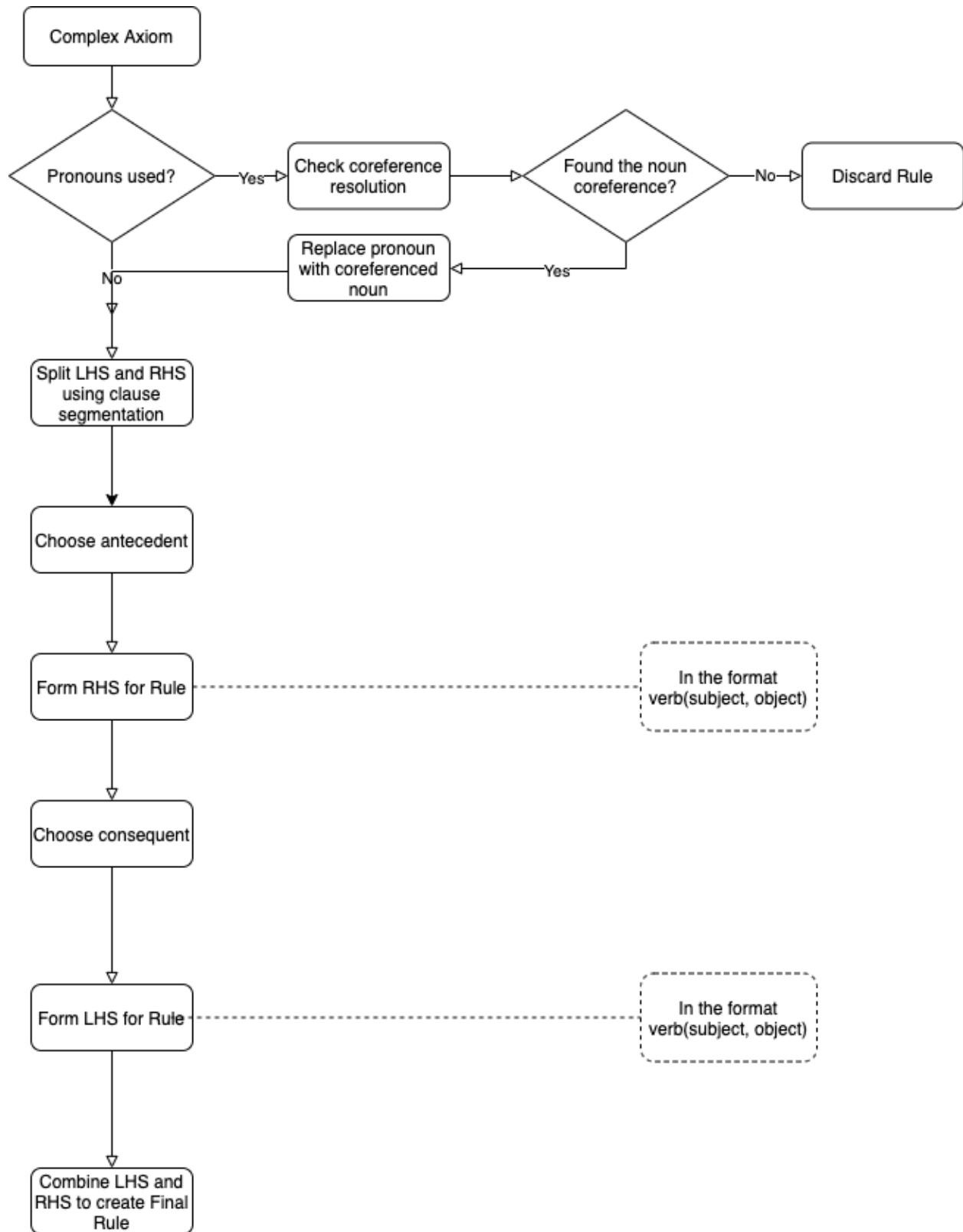


Figure 5.3: Processing complex axioms involves separating out the antecedent and consequent, identifying the concepts mentioned in those parts, and storing those concepts in the rule base.

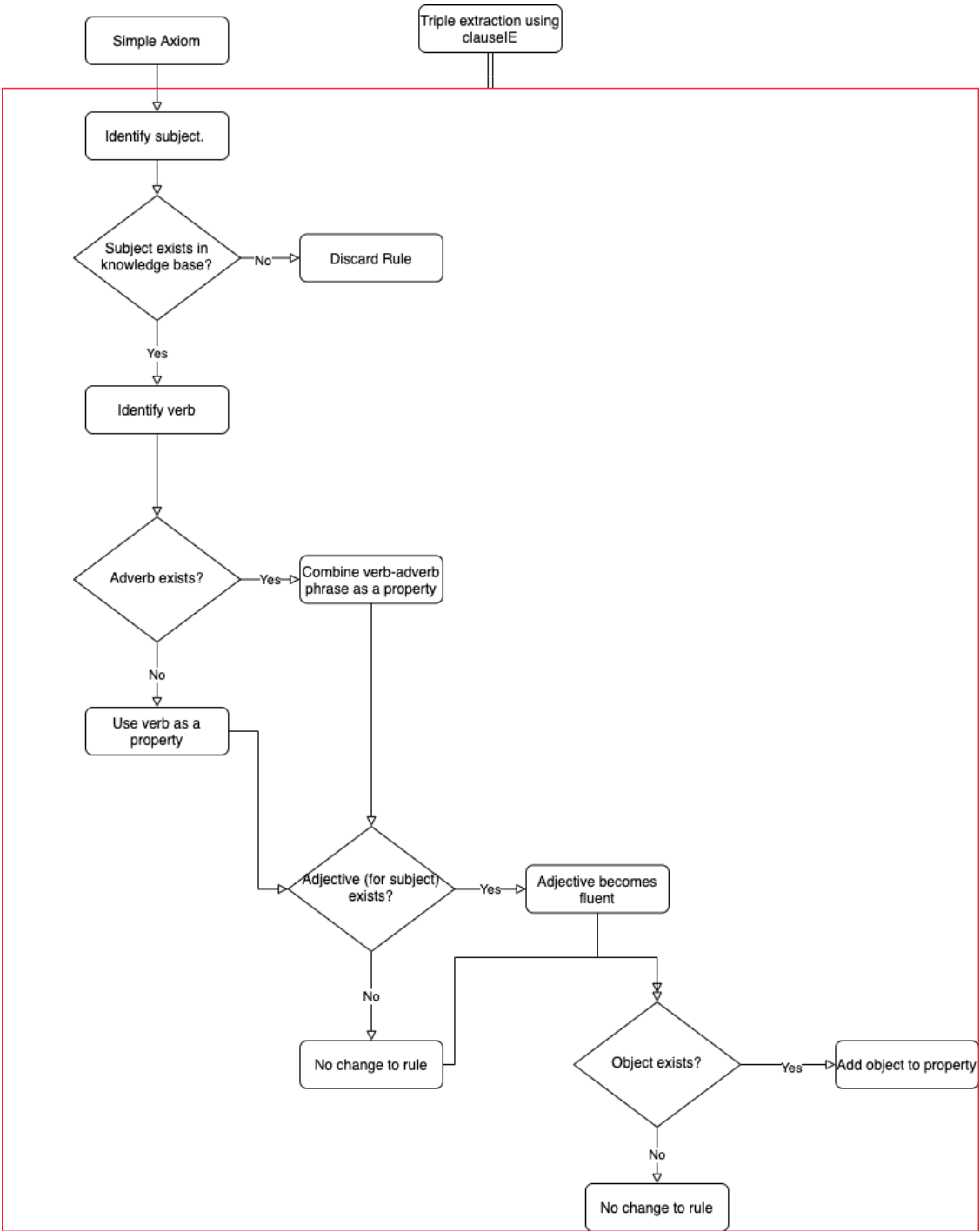


Figure 5.4: Processing simple axioms broadly involves identifying the subject, action(verb) and object. We extract triples using the prepositions identified by the python implementation of ‘clauseIE’.

CHAPTER 6

EVALUATING RULES EXTRACTED FROM A TEXT NARRATIVE

6.1 Need for Evaluation

”A key factor which makes a particular discipline or approach scientific is the ability to evaluate and compare the ideas within the area” [184]. Thus, when exploring new research areas it is essential to have a method to evaluate the success of the experiment.

The main contribution of this thesis is the semi-automated generation of rules using neural networks. Automated Axiom generation tasks do not have an established method of evaluating the results produced by the method where the rules are not explicitly stated in the document. In this chapter, we explore the existing research in the fields of ontology evaluation, and expert system evaluation to identify the important factors in evaluation of a fact base and more importantly the rule base (which in our method has been generated using crowdsourcing and Text2Rule conversion, see chapters 3 and 5) of a GOF AI system.

6.2 What Can We Learn from Existing Evaluation Approaches?

6.2.1 Ontology Evaluation

”Ontology evaluation is the task of measuring the quality of an ontology” [185]. ”Most evaluation approaches in ontologies fall into one of the following categories” [184]:

- Comparing the ontology to a ”golden standard” [184],[186].
- Utilizing the ontology as part of an application and evaluating those results [184],[187].
- Comparing with a source of data about the domain documented in the ontology [184],[188].
- Having humans evaluate the ontology ”based on a set of predefined criteria, standards and requirements” [184],[189].

When we broadly compare ontology development and axiom generation and try to identify what category of approaches would be most effective at evaluating automatically generated axioms, human evaluation seems like the optimal option. Since writing rules has

always been a task performed by human experts, it would ideal to have humans evaluate the rules of a GOF AI system. When considering human evaluation, the most important factor in constructing an evaluation approach is to choose the right criteria using which an expert can evaluate the results (in our case, the axioms).

Burton-Jones et. al [190], describes a set of semiotic metrics to assess the quality of an ontology. These metrics were designed to "assess the syntactic, semantic, pragmatic, and social aspects of ontology quality" [190]. The metrics introduced by the authors aren't necessarily all human evaluated. Some metrics used are evaluated using data driven methods. The rating criteria used by authors are [190]:

- Lawfulness : Focuses on identifying syntactic errors in the ontology.
- Richness : Based on the number of the properties used to describe the ontology.
- Interpretability : Measure by checking if the terms in the ontology are meaningful.
- Clarity : An extension of the interpretability metric, clarity checks for the number of senses in WordNet. The goal of having high clarity is for terms in the ontology not to be ambiguous.
- Consistency : This metric focuses on the internal inconsistencies in the ontology. Inconsistencies in an ontology refers to when having incoherent logical inferences, such as contradictions in the ontology.
- Comprehensiveness : Counts the number of classes and properties in an ontology.
- Accuracy : Measure of the accuracy of the knowledge given by the ontology. Accuracy is measured by domain experts, when they test their knowledge against the inferences output by the ontology.
- Authority : This metric is measured based on the number of reference made to an ontology from other ontologies.
- Relevance : This metric examines the degree to which the ontology provides information of the type needed by an application.
- History : Focuses tracking usage of an ontology by other applications.

6.2.2 Expert System Evaluation

As seen in section 2.2.2, expert systems consist of knowledge bases, rule bases and a set of termination criteria. The amount of evaluation performed on an expert system and the importance attached to it, depends on the size, complexity, criticality and other aspects of the expert system [191]. Grogono et. al presents an approach to verify a knowledge base, including the rules of an expert system. The authors do this by detecting anomalies in the construction of the knowledge base. Anomalies are not necessarily errors, but they indicate flaws in the system which may lead to errors. The four kinds of anomalies described are [191]:

- Ambivalence : When the inferred set of hypothesis contains semantic constraints, the knowledge base is considered ambivalent. Ambivalence may be caused by human experts holding different views and/or mistakes in the rules entered.
- Circularity : If rules are continuously fired in an infinite loop, the knowledge base is circular.
- Redundancy : If the inclusion or omission of a literal or a rule make no difference to inference made by the AI system, then the knowledge base contains redundancies.
- Deficiency : When a knowledge base is unable to infer a final hypothesis for a specific environment, even though according to the system specifications it should reach a final goal.

6.3 Are They Directly Applicable?

Since generating axioms is the main focus of the experiments mentioned in this thesis, the evaluation metrics chosen should be specifically focusing on reviewing the various aspects of an axiom. An axiom or a rule in a Symbolic AI system typically consists of antecedents and consequents, but for this purpose we are interested in evaluating the rules generated and hence must confirm the soundness of the antecedent and consequents statements themselves. Hence the granular components of rules may be listed as :

- Concepts : Concepts may refer to instances of actors, objects or entities that may be relevant, and present in the knowledge base. This generally holds true until the rule calls for the creation of a new instance

- **Conditions** : Most rules in expert systems follow the IF-THEN format, but there are other conditions that can be seen used in AI systems, like *iff* or *while* conditions. The antecedent component of the rule typically contains conditions, since it is the execution of the condition that leads to the firing of the rules.
- **Actions** : Agents or Actors perform actions. Execution of actions takes place in the antecedents and the consequents and is essential temporal ordering of events and situations. In Situation Calculus, action play an essential role, since according levesque et. al [45], "a situation is a history or a finite sequence of actions".

6.3.1 Adapting Metrics

In this chapter we develop metrics to evaluate the axioms generated by our proposed method. To do this, we must first evaluate components of the axioms stated above. Some of the metrics stated in section 6.2, may not be directly applied to axioms, but can be modified to do so.

For example, the **Intrepretability** metric would measure if the concepts used in the rules are meaningful and/or exist in the knowledge base. For text generation tasks, this would be a particularly important and challenging metric (in terms to achieving a high score), because character level models are known to generate different words from those that exist in the corpus. Measuring this metric can be automated, since comparison with the knowledge base can done automatically using any commonly used pattern matching algorithm. And we can compare the concepts to a dataset like WordNet [192] to find out if the words are meaningful, similar to the method described in [190]. We define interpretability as :

$$Interpretability = \frac{Count(Concepts_in_RuleBase \cap Concepts_in_KnowledgeBase)}{Count(Concepts_in_RuleBase)} \quad (6.1)$$

Lawfulness measures the syntactic errors in the rulebase. Since axioms are meant to be machine readable any syntactic errors should be easily identifiable. In addition to errors generated by the system, we will also check for anomalies such as those mentioned in section 6.2.2 [191]. While identification of errors maybe automated, anomaly identification needs to be done by human experts.

Accuracy measures if the rules generated conform to domain knowledge. This metric also needs to be evaluated by a group of domain expert can examine the rules and test the system for logical inferences. This metric has 2 steps in the evaluation, (1) First we establish if the rule needs a domain expert to evaluate the accuracy. If the rule does not require knowledge of domain (i.e the need to read the story), we classify it under the category $Accuracy_C$, which implies that the resultant statement/condition is a commonsense rule. (2) If the human evaluation requires knowledge of the domain, then statement/condition is classified under the category $Accuracy_A$, for all other rules. If NR is the total number of rules, and T is number of rules that are true according to human evaluators, $Accuracy_A$ is measured as $\frac{T}{NR}$. $Accuracy_C$ is measured with the same formula, but only considering common sense rules as the complete set.

6.4 New Metric - Coverage

One of the metrics that is essential to evaluate generated axioms is **coverage**. This metric does not evaluate one rule at a time, instead it is evaluated over an automatically generated rule base. Coverage measures how well the rule base covers the knowledge of the selected world to be modeled. Coverage and accuracy have differences in what they measure. While accuracy measure whether a rule matches the knowledge of the domain, the coverage metric measures whether a rule base covers the known laws in the domain.

Coverage is calculated using the TF-IDF (Term Frequency - Inverse Document Frequency) approach [193] at the fact level. In our approach to calculate the coverage of a rule base we call the TF equivalent as Fact Frequency (FF) and the IDF equivalent is called Coverage Constant (CC). If a fact occurs in the rule base and the fact bases, the FF-CC score will always be 0. If a fact does not occur in either the rule base or the fact base, the FF-CC score will be greater than 0. Summation of the FF-CC score for all facts in the rule base is the coverage score. The coverage score can be normalized by dividing the obtained score from the previous step by the count of the facts commonly occurring in the knowledge base and the rule base. This normalization is done in the case where a comparison of multiple knowledge and rule bases is being conducted with a significant difference in the number of

facts present in the data being compared. Ideal coverage score for a rule base is 0, lower the coverage score, the better the coverage of the rule base.

$$FF = Count(Fact_{KB}) \quad (6.2)$$

$$CC = \begin{cases} 1, & \text{if the fact occurs both in the knowledge base and rule base} \\ 0.5, & \text{if the fact occurs only in the knowledge base} \end{cases} \quad (6.3)$$

$$Coverage = \sum FF * \log(CC) \quad (6.4)$$

$$Coverage_{normalized} = \sum \frac{FF * \log(CC)}{NF_{RB}} \quad (6.5)$$

,where NF_{RB} = Total number of facts in the rule base.

For example, if a knowledge base has 20 facts and the corresponding rule base has 8 rules. We then examine the facts mentioned in both the knowledge base and rule base. Each fact in the knowledge base is checked to see if it is called in the rule base. If a fact that appears 3 times in the knowledge base also occurs in the rule base then:

$$FF = 3 \quad (6.6)$$

$$CC = 2/2 = 1 \quad (6.7)$$

$$FF - CC = 3 * \log(1) = 0 \quad (6.8)$$

Similarly, we sum up all the FF-CC scores to form the final Coverage score.

6.5 Results

The final results of the thesis workflow was obtained after the execution of the text to rules contribution. As mentioned in Chapter 3 and 5, the results produced during crowd-sourcing were inconsistent and thus unfit for processing in an automated computational workflow. 3 out of the 4 evaluation metrics/methods have been implemented on the results

obtained from running the entire SAAG workflow and these findings can be found in Appendix F. In this section we describe the findings of our evaluation method and observations on studying those findings.

6.5.1 Interpretability

Interpretability measures if concepts used in the rules are meaningful and/or exist in the knowledge base. The formula to calculating the interpretability of a rule base is defined in Section 6.3.1. As mentioned in Sections 5.4 and 4.2, we can see that one of limitations of the rules produced from crowdsourcing were the inconsistent references to the same entity using different names. This inconsistency has been highlight and even quantified during the execution of the interpretability calculation. The highest interpretability score obtained for a rule base was 0.214, and the lowest interpretability score was 0. Which means that the best "knowledge base - rule base" corpus created using our SAAG workflow had 21.4% interpretability and the worst "knowledge base - rule base" corpus had no concepts in common (thus the interpretability score of 0), which meant that no rules would be executed during the querying of the knowledge base. The average interpretability score of all the corpora produced using the SAAG workflow was 0.067. Complete results of the interpretability scores for all corpora can be found in Appendix F.1.

6.5.2 Coverage

Our approach to calculating the coverage of a rule base used a counter intuitive application of the TF-IDF metric [193]. Calculating the coverage of a rule base involved a fact-level examination of both the knowledge base and the rule base. Since facts are made up of concepts, and we quantified the inconsistencies in the concepts during the interpretability evaluation, we those inconsistencies are not only evident but also enhanced at the fact-level examination of the rule base. Since we were comparing "knowledge base - rule base" corpora of significantly varying sizes, we had to normalize the coverage to examine the rule bases on the same scale. The normalized coverage showed there were only 2 rule bases had facts that covered some portion of the knowledge base. Most of the normalized coverage scores were $-\infty$, which meant that none of the rules in the rule base covered the knowledge base at all.

An interesting and important point observed by examining the results was that rule bases with a relatively high interpretability score can still have very low coverage. This is

because the interpretability is examined at the concept level, while coverage is calculated at the fact level. And at the end of the day, Golog or other logical representation languages view a fact as the representation of a datum. Complete coverage and normalized coverage scores can be found in Appendix F.1.

6.5.3 Lawfulness

Lawfulness was calculated simply by the count of syntactic errors in the rule base. In order to evaluate the lawfulness, we simply loaded the rule bases into the Golog interpreter built in SWI-prolog¹³. There were 204 errors produced, which means that 930 lawful rules generated in 65 rule bases by implementing the SAAG workflow. Since our current iteration of the SAAG workflow relied on known NLP methods to extract concepts to create facts and rules, the number of errors produced in creating the rule bases was low. The detailed error outputs on loading the rule bases can be found in Appendix F.2.

6.6 Future Work

The evaluation method described in this chapter was created to be applied to all "knowledge base - rule base" corpora, not just to the results of the SAAG workflow. The metrics devised for this evaluation can be applied to corpora compiled in various languages used to build symbolic AI systems. One of the key direction for the future future work of this evaluation method involves applying the evaluation metrics to corpora from different domains and encoded in different languages. Another metric that needs to be completely tested is Accuracy.

6.6.1 Accuracy

Accuracy measures if the rules generated for the rule base conform to domain knowledge. Of all the metrics developed for this contribution, Accuracy is the only metric that is applicable to automatically generated rule bases. If rule bases are created by experts, then the rules are going to be an accurate representation of the domain. As mentioned in section 6.3.1, accuracy is calculated at 2 venues: one that checks if the rule needs a domain expert to evaluate the accuracy of the rule, there by labeling the common sense (CSR) and non commense rules (NCSR) and second that actually has domain experts evaluate the NCSRs.

¹³http://www.cs.toronto.edu/cogrobo/Systems/golog_swi.pl

In this thesis, we have completed the first step and successfully labeled CSRs and NCSRs. The second step though requires an independent evaluation conducted by domain experts. This step is time consuming and needs a set of domain experts to complete the process in a timely manner. Such an endeavour may cost more money than the original crowdsourcing experiment, and as a result is out of the scope of this thesis.

Also accuracy as metric makes more intuitive sense when the rules are generated in a fully automated manner such as deep learning method used to generate rules for a given knowledge base (see section 7.4.3). Thus we plan to apply this metric along with all the other evaluation metrics to the results produced by our exploration detailed in the long terms future plans section in chapter 7 of this thesis document.

6.7 Testing Wider Applicability

A long term vision for this evaluation method is to apply these metrics to many "knowledge base - rule base" corpora. A major point of focus for the future will involve testing the evaluation metrics on many datasets to show the value of evaluating rule bases this way. We approach this goal by evaluating many existing rule rich ontologies, expert systems and other symbolic AI applications using our metrics. We also plan to refine these metrics as we learn of inadequacies in their formulation during the testing phase.

With the increase in the automation of various parts of expert system design pipeline [22],[28],[29],[108],[109], our evaluation method becomes an integral part of the future landscape of rule based AI.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Thesis Review

Automated Axiom generation is a difficult research problem to solve. Most of the work in this area focuses on extracting rules from text documents with formal language, that explicitly mention rules in the text [22]–[29]. This thesis developed a semi-automated method to generate axioms for a set of text narratives, when rules aren't explicitly mentioned in the text.

Chapter 1 introduces the research problem we explore during this thesis, the research methodology developed to solve this research problem, the key contributions that emerged in the execution of this methodology, and broader impact of this work. Chapters 3,4, and 5 focus on the individual steps of the thesis methodology.

Chapter 2 takes an in depth look at the history of Artificial Intelligence (AI). We specifically focus describing the evolution of Rule based AI, and the field of Natural Language Processing (NLP), and the current state of the field. Understanding the history of these fields helps identify the white spaces and unanswered scientific questions that continue to loom over the researchers. This work provided the motivation in creating the larger research direction for this thesis, and helped identify how certain techniques are useful to execute specific parts of the Semi-Automated Axioms Generation (SAAG) workflow.

In **chapter 3**, we create rules in natural language based a narrative using amazon mechanical turk, a popular crowdsourcing platform. This chapter also explores the data used for the crowdsourcing experiments, the limitations to this approach and how we overcome these limitations. Also explored are the incentives for the participants in the experiment, along with curation and evaluation criteria that are used for our experiment.

The results of the crowdsourcing experiment are inconsistent and cannot be used in a workflow where these results need to be processed automatically as part of larger computational methodology. Hence, in **chapter 4**, we document the results from chapter 3, point out the issues with these results, and how we overcome these inconsistencies with an human evaluation of crowdsourcing results. We also expand on the lessons learnt from the crowdsourcing experiment, by observing and exploring the process in the context of expert replacements. This exploration led to us finding a set of boundaries for the usability of

crowdsourcing as tool or means to overcome the automation bottleneck where human intelligence is required for certain steps in the scientific workflow. We also discuss in detail the concept of "Humans in the loop" (HITL), and "Experts in the loop" (EITL) and how they are distinct from each other. We provide a platform for further exploring and developing a framework for EITL workflows, and document future directions this work that will aid in the workflow design process for artificial intelligence applications.

In **chapter 5**, we converted the results from the crowdsourcing into formal situation calculus with a workflow developed using known NLP techniques. We also examined the results from the text to rule conversion method and pointed out the limitations of using only NLP techniques to extract formal logical rules from raw text. Additionally, we propose a fully automated method based on machine translation applications to convert text to formal rules.

Chapter 6 presents an method to evaluate "knowledge base - rule base" corpora for the usage in symbolic AI applications. In this chapter we adapted some existing metrics from ontology and expert system evaluation methods and introduced a novel metric called 'coverage' to measures how well a rule base covers the knowledge represented in the world being modeled. We also describe the results from the evaluation method when applied to the rule generated from the SAAG workflow.

7.2 Lessons Learnt

This thesis started as an ambitious exploration of the boundaries of artificial intelligence. After observing the evolution of the fields of machine learning and symbolic AI in parallel and the rarity in the efforts to combine them, we developed an approach to automatically generate rules using deep learning for a given knowledge base in any domain. However, the feasibility and validity of this method could not be explored or tested because of the lack of "knowledge base - rule base" corpora available for training. Thus, there was a shift in focus to developing a method to generate these corpora without needing pre-compiled and curated training datasets.

Our proposed workflow for this new exploration was a semi-automated axiom generation (SAAG) workflow which used crowdsourcing as the key step to generate these rules. Observing the results of the crowdsourcing method and the contributions that followed, we understood the boundaries of usability of crowdsourcing for tasks where crowd participation

is considered for replacing experts in the workflow. We also found that crowdsourcing and the inconsistencies caused by the results of **complex** crowdsourcing tasks, make it very difficult for it to be included in automation driven pipelines or even computational workflows in general. Since crowdsourcing was envisioned to be the cornerstone of the SAAG workflow, the effect of crowdsourcing results ripple down the next steps of the workflow.

This exploration however, sparked new ideas for improving the individual contributions of this thesis and thus be better equipped for delving head first into the extremely complex and still unsolved problem of fully automated axiom generation. In the next section, we expand upon our future plans to continue, improve and expand on our work on automated axiom generation, so that we can explore new venues and face new hurdles in the quest to answer the question "To what extent can we fully automate the axiom generation process for GOFAI?"

7.3 Impact and Future Applications

The SAAG workflow presented in this thesis has produced a set of rules bases and fact bases generated from 65 project gutenburg short stories. The impact of this work though, can be seen in more than just a directly application or reproduction of the workflow. The "natural language rule dataset" produced as an intermediate step of the SAAG workflow can be used as a summary of the world shown in the story. This dataset may be used in abstractive [194],[195] or extractive summarization [196] training tasks (by combining both the stories and the text rules), or as a starting point for a textual entailment [197] dataset (by just using the rules).

In addition to the data produced from this thesis, the metrics developed and observations made in this thesis will improve scientific research not only in artificial intelligence, but in the many interdisciplinary scientific explorations. The evaluation metrics presented in chapter 6 can be used to evaluate any "knowledge base-rule base" combination for any logical language and any domain. Our examination of the crowdsourcing results showed the boundaries of crowdsourcing as an approach and its use in automated workflows. We also showed the need for the term Experts in the loop (EITL) to be distinct from the Humans in the loop. This is because there are distinct considerations and factors to include "experts in the loop" vs "non-experts in the loop". Part of the future work of this thesis is to create a framework for experts in the loop for scientific workflows. The impact of such a

framework can be felt in "workflow design" (for both workflow systems or just for scientific computational workflow) in the future. Citizen science initiatives and expert driven scientific explorations will also benefit from understanding better when and how human and expert involvement in scientific workflows will improve their research. In its entirety, this thesis work can also be considered a stress test of certain facets of artificial intelligence. The detailed documentation of the white-spaces and limitations of commonly used approaches in AI provide much needed direction for future development of AI methods and novel applications of those methods.

7.4 Future Work

Scientific explorations that navigate through uncharted territories always produce a array of interesting research paths, alternative approaches and improvements on the current approach (both planned and unplanned). This thesis is no different, and has produced ideas for future research and explorations (both short term and long term) that will shape the future of AI, and will help think about Symbolic AI and Machine learning in a more symbiotic environment, where advances in one area can and will influence the evolution of the other. In this section, we document some of the future researched planned for the field of Automated Axiom Generation (AAG).

7.4.1 Short Term

The Sentence Ordering LSTM model (See chapter 2) was chosen based on the a detailed study that compared all the common text embedding models to predict the order of the sentences in a paragraph [70]. But the study has not compared some methods that are very popularly used for deep learning on image data, like Generative Adversarial Networks (GANs), and Capsule Network (CapsNet). One of the short terms plans for the future is to test out how the sentence ordering model can be improved by testing both GANs and CapsNet models and see how they compare to LSTM networks.

A deep learning method for converting text in natural language to formal rules (Chapter 5) was being developed using machine translation as a the basis for this conversion. But this exploration could not be tested because there was no data available to train the neural network model. In the future, we plan to use the dataset compiled during this thesis and

datasets we can collect from contacting other researchers in the field to train a deep learning model to convert text from natural language to formal rules in situation calculus.

7.4.2 Mid Term

The contributions in chapter 3 led to some interesting discoveries about crowdsourcing experiments and about Humans in the loop workflows in general. With this work we defined some boundaries on the applicability of crowdsourcing, and its use in the context of expert replacements. We also documented the distinction between "Humans in the loop" and "Experts in the loop" workflows. Future work for this exploration focuses around comparing HITL and EITL workflows and creating a complete framework for "Experts in the loop" in artificial intelligence, including documenting the boundaries of the approach, the advantages, limitations and decision making criteria.

7.4.3 Long Term

The original work being explored during this thesis was the development of an automated axiom generation methodology, i.e. learning to write axioms for a given knowledge base. Automating the axiom generation process has been proven to be a difficult research question to solve. One long term goal from this thesis is to explore the area of automated axiom generation by looking at the intersection of the symbolic AI and machine learning, and see if a machine can write axioms about any given domain. One such method involves creating a deep learning model that is trained on existing situation calculus datasets (both the factbases and rulebases), with new rules as a desired output. The goal of the deep neural networks would be to find the mapping between the concepts in the factbase and the required axioms. Once the model produces an initial set of axioms/rules, the model will be iteratively improved using reinforcement learning. In the reinforcement process, each rule will be evaluated and assigned a score, if the score is below the predefined threshold, then the rule is rejected.

This is but one of the many potential methods to create an automated axiom generation workflow, the majority of which have not been explored yet. The long-term goal of this exploration is to develop a system which can taken in facts about any given domain and accurately state the rules for that domain. To achieve this, detailed experiments and exploration much be done on transferability of the concepts learned from one domain to another

completely unrelated domain. Quantifying transferability between domains, and scalability is another goal. For e.g. Can we use small training corpora and test on large ones, or to see if a demo, like the AllenNLP [198] semantic parsing demo¹⁴ is possible?

¹⁴<http://demo.allennlp.org/nlvr-parser/>

REFERENCES

- [1] P. McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natick, MA, USA: AK Peters Ltd, 2004.
- [2] J. McCarthy, “What is artificial intelligence?” Stanford Univ., Stanford, CA, USA, Tech. Rep., 2007. Available: <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf> Accessed: July 29, 2021
- [3] N. Nilsson, *The Quest for Artificial Intelligence : A History of Ideas and Achievements*. Cambridge, UK: Cambridge Univ. Press, 2009.
- [4] R. C. Schank, “What is ai, anyway?” *AI Mag.*, vol. 8, no. 4, pp. 59–59, Dec. 1987.
- [5] P. Maes, “Behavior-based artificial intelligence,” in *Proc. of the 15th Annu. Meeting of the Cogn. Sci. Soc.*, 1993, pp. 74–83.
- [6] R. R. Trippi and E. Turban, *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. New York, NY, USA: McGraw-Hill, Inc., 1992.
- [7] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, “Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer.” *Radiol.*, vol. 187, no. 1, pp. 81–87, Apr. 1993.
- [8] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. New York, NY, USA: Pearson, 2016.
- [9] J.-C. Pomerol, “Artificial intelligence and human decision making,” *Eur. J. of Oper. Res.*, vol. 99, no. 1, pp. 3–25, May 1997.
- [10] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM J. of Res. and Develop.*, vol. 3, no. 3, pp. 210–229, Jul. 1959.
- [11] J. Weizenbaum, “Eliza - a computer program for the study of natural language communication between man and machine,” *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966.
- [12] W. R. Swartout, “Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: B.G. Buchanan and E.H. Shortliffe, (addison-wesley, reading, ma, 1984); 702 pages, \$40.50,” *Artif. Intell.*, vol. 26, no. 3, pp. 364 – 366, Jul. 1985. Available: <https://www.sciencedirect.com/science/article/abs/pii/0004370285900670> Accessed: July 31, 2021
- [13] F.-H. Hsu, *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton, NJ, USA: Princeton Univ. Press, 2002.

- [14] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, and et al., “Mastering the game of go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, Oct 2017.
- [15] J. Haugeland, *Artificial Intelligence: The Very Idea*. Cambridge, MA, USA: MIT Press, 1989.
- [16] T. Hobbes, *Leviathan*. Lexington, KY: Seven Treasures Publications, 2009.
- [17] C. M. Keet, “Open world assumption,” in *Encyclopedia of Syst. Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 1567–1567.
- [18] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.
- [19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [20] J. Bachant and J. McDermott, “R1 revisited: Four years in the trenches,” *AI Mag.*, vol. 5, no. 3, pp. 21–32, Sep. 1984.
- [21] C. Grosan and A. Abraham, *Intelligent Systems: A Modern Approach*, vol. 17. Cham, Switzerland: Springer, 2011.
- [22] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori, “Combining natural language processing approaches for rule extraction from legal documents,” in *Lecture Notes in Comput. Sci.* Cham, Switzerland: Springer Int. Publishing, Oct. 2018, pp. 287–300.
- [23] E. Francesconi, “Legal rules learning based on a semantic model for legislation,” in *Proc. of the LREC 2010 Workshop on the Semantic Process. of Legal Texts (SPLeT-2010)*, Malta, May 2010, p. 46.
- [24] A. Z. Wyner and W. Peters, “On rule extraction from regulations,” in *Frontiers in Artif. Intell. and Applications*, vol. 235. Amsterdam, Netherlands: IOS Press, 2011, pp. 113–122.
- [25] A. Wyner, T. Van Engers, and K. Bahreini, “From policy-making statements to first-order logic,” in *Int. Conf. on Electron. Govt. and the Inf. Syst. Perspective*. Springer, 2010, pp. 47–61.
- [26] A. Boufrida and Z. Boufaida, “Automatic rules extraction from medical texts,” in *2014 Int. Workshop on Adv. Inf. Syst. for Enterprises*. Tunis, Tunisia: IEEE, Nov. 2014, pp. 29–33.
- [27] J. Delannoy, C. Feng, S. Matwin, and S. Szpakowicz, “Knowledge extraction from text: Machine learning for text-to-rule translation,” in *Proc. of Eur. Conf. on Mach. Learn. Workshop on Mach. Learn. and Text Anal.*, 1993, pp. 1–7.

- [28] D. Kholkar, S. Sunkle, and V. Kulkarni, "Semi-automated creation of regulation rule bases using generic template-driven rule extraction," in *Proc. of the 2nd Workshop on Autom. Semantic Anal. of Inf. in Legal Text*, vol. 2143. London, UK: CEUR Workshop Proc., Jun. 2017, paper 1.
- [29] K. Fatema, C. Debruyne, D. Lewis, D. O'Sullivan, J. P. Morrison, and A. A. Mazed, "A semi-automated methodology for extracting access control rules from the european data protection directive," in *2016 IEEE Secur. and Privacy Workshops*. IEEE, May 2016, pp. 25–32.
- [30] S. Lahiri, "Complexity of word collocation networks: A preliminary structural analysis," in *Proc. of the Student Res. Workshop at the 14th Conf. of the Eur. Chapter of the Assoc. for Comput. Linguistics*. Gothenburg, Sweden: Assoc. for Comput. Linguistics, Apr. 2014, pp. 96–105.
- [31] F. Lin, "Situation calculus," in *Handbook of Knowl. Representation*, vol. 3, ser. Found. of Artif. Intell., F. van Harmelen, V. Lifschitz, and B. Porter, Eds. Amsterdam, Netherlands: Elsevier, 2008, pp. 649–669.
- [32] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," 2015, *arXiv:1509.00685*.
- [33] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," 2018, *arxiv:1812.02303*.
- [34] H. Liu, A. Gegov, and M. Cocea, *Rule Based Systems for Big Data*. Cham, Switzerland: Springer Int. Publishing, 2015.
- [35] J. McDermott, "R1: A rule-based configurer of computer systems," *Artif. Intell.*, vol. 19, no. 1, pp. 39–88, Sep. 1982.
- [36] B. Buchanan, *Rule-based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA, USA: Addison-Wesley, 1984.
- [37] J. Gaschnig, "Prospector: An expert system for mineral exploration," in *Introductory Readings in Expert Systems*, D. Michie, Ed. New York, NY, USA: Gordon and Breach Sci. Publishers, 1982, ch. 3, pp. 47–64.
- [38] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "Dendral: a case study of the first expert system for scientific hypothesis formation," *Artif. Intell.*, vol. 61, no. 2, pp. 209–261, Jun. 1993.
- [39] A. M. Lumb, R. B. McCammon, J. L. Kittle *et al.*, *Users Manual for an Expert System (HSPEXP) for Calibration of the Hydrological Simulation Program–Fortran*. Reston, VA, USA: US Geological Survey, 1994.
- [40] F. Hayes-Roth, "Rule-based systems," *Commun. of the ACM*, vol. 28, no. 9, pp. 921–932, Sep. 1985.

- [41] A. Abraham, “Rule based expert systems,” in *Handbook of Measuring System Design*. Chichester, UK: Wiley, 2005, ch. 130.
- [42] J. McCarthy, “Situations, actions, and causal laws,” Stanford Univ., Stanford, CA, USA, Tech. Rep. AD0785031, Jul. 1963.
- [43] R. Reiter, “The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression,” *Artif. Intell. and Math. Theor. of Comput. : Papers in Honor of John McCarthy*, vol. 27, pp. 359–380, Sep. 1991.
- [44] J. McCarthy and P. J. Hayes, “Some philosophical problems from the standpoint of artificial intelligence,” in *Machine Intelligence 4*, B. Meltzer and D. Michie, Eds. Edinburgh, UK: Edinburgh Univ. Press, 1969, pp. 463–502.
- [45] H. Levesque, F. Pirri, and R. Reiter, “Foundations for the situation calculus,” in *Linköping Articles in Computer and Information Science*, vol. 3, E. Sandwell, Ed. Linköping, Sweden: Linköping Univ. Electron. Press, 1998.
- [46] J. McCarthy and T. Costello, “Combining narratives,” in *Proc. of the 6th Int. Conf. on Princ. of Knowl. Representation and Reasoning*, vol. 98. Trento, Italy: Morgan Kaufmann, Jun. 1998, pp. 48–59.
- [47] J. D. Funge, “Making them behave: Cognitive models for computer animation,” Ph.D. dissertation, Univ. of Toronto, Toronto, Canada, 1998.
- [48] H. J. Levesque, R. Reiter, Y. Lespérance, F. Lin, and R. B. Scherl, “GOLOG: A logic programming language for dynamic domains,” *The J. of Log. Program.*, vol. 31, no. 1, pp. 59–83, Jun. 1997.
- [49] P. A. Bonatti and D. Olmedilla, “Rule-based policy representation and reasoning for the semantic web,” in *Reasoning Web*. Berlin, Germany: Springer, 2007, pp. 240–268.
- [50] G. De Giacomo, Y. Lesprance, and H. J. Levesque, “ConGolog, a concurrent programming language based on the situation calculus,” *Artif. Intell.*, vol. 121, no. 1, pp. 109–169, Aug. 2000.
- [51] L. Deng and D. Yu, “Deep learning: Methods and applications,” Microsoft, Seattle, WA, USA, Tech. Rep. MSR-TR-2014-21, May 2014. Available: <https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/> Accessed: July 30, 2021
- [52] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bull. of Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.
- [53] K. J. Cios, “Deep neural networks - A brief history,” 2017, *arXiv:1701.05549*.
- [54] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC, USA: Spartan Books, 1962.

- [55] K. Gurney, *An Introduction to Neural Networks*. London, UK: CRC Press, 2014.
- [56] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969.
- [57] M. L. Minsky, “Theory of neural-analog reinforcement systems and its application to the brain-model problem,” Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1954.
- [58] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. of the Nat. Acad. of Sci.*, vol. 79, no. 8, pp. 2554–2558, Apr. 1982.
- [59] P. Werbos, “Beyond regression: New tools for prediction and analysis in the behavioral sciences,” Ph.D. dissertation, Harvard Univ., Cambridge, MA, USA, 1974.
- [60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct 1986.
- [61] P. Langley, “The changing science of machine learning,” *Mach. Learn.*, vol. 82, no. 3, pp. 275–279, Mar 2011.
- [62] C. Baral, O. Fuentes, and V. Kreinovich, “Why deep neural networks: A possible theoretical explanation,” in *Constraint Programming and Decision Making: Theory and Applications*, M. Ceberio and V. Kreinovich, Eds. Cham, Switzerland: Springer, 2018, pp. 1–5.
- [63] X. Ying, “An overview of overfitting and its solutions,” *J. of Phys.: Conf. Ser.*, vol. 1168, no. 2, Feb. 2019, Art. no. 022022.
- [64] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, “Recurrent neural network based language model,” in *11th Annu. Conf. of the Int. Speech Commun. Assoc.*, Makuhari, Japan, Sep. 2010, pp. 1045–1048.
- [65] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [66] R. Barzilay, N. Elhadad, and K. R. McKeown, “Sentence ordering in multidocument summarization,” in *Proc. of the 1st Int. Conf. on Human Lang. Technol. Res.* Assoc. for Comput. Linguistics, Mar. 2001, pp. 1–7.
- [67] N. Okazaki, Y. Matsuo, and M. Ishizuka, “Improving chronological sentence ordering by precedence relation,” in *Proc. of the 20th Int. Conf. on Comput. Linguistics*. Geneva, Switzerland: Assoc. for Comput. Linguistics, Aug. 2004, pp. 750–es.
- [68] D. Bollegala, N. Okazaki, and M. Ishizuka, “A bottom-up approach to sentence ordering for multi-document summarization,” *Inf. Process. & Manag.*, vol. 46, no. 1, pp. 89–109, Jan. 2010.

- [69] P. Li, G. Deng, and Q. Zhu, “Using context inference to improve sentence ordering for multi-document summarization,” in *Proc. of 5th Int. Joint Conf. on Natural Lang. Process.*, Chiang Mai, Thailand, Nov. 2011, pp. 1055–1061.
- [70] X. Chen, X. Qiu, and X. Huang, “Neural sentence ordering,” 2016, *arXiv:1607.06952*.
- [71] S. Joty, G. Carenini, and R. T. Ng, “Topic segmentation and labeling in asynchronous conversations,” *J. of Artif. Intell. Res.*, vol. 47, no. 1, pp. 521–573, May 2013.
- [72] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, “Topical clustering of tweets,” in *Proc. of the Assoc. for Comput. Mach. Workshop on Social Web Search and Mining*, no. 63, Beijing, China, Jul. 2011.
- [73] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, “Topical word embeddings,” in *Proc. of the 29th AAAI Conf. on Artif. Intell.*, Austin, TX, USA, Jan. 2015, pp. 2418–2424.
- [74] J. H. Lau and T. Baldwin, “An empirical evaluation of doc2vec with practical insights into document embedding generation,” 2016, *arXiv:1607.05368*.
- [75] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, *arXiv:1301.3781*.
- [76] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *7th Int. Conf. on Document Anal. and Recognit.* Edinburgh, UK: IEEE, Aug. 2003, pp. 958–963.
- [77] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” 2017, *arXiv:1710.09829*.
- [78] K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” 2014, *arXiv:1406.1078*.
- [79] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” 2014, *arXiv:1412.6632*.
- [80] I. Sutskever, J. Martens, and G. E. Hinton, “Generating text with recurrent neural networks,” in *Proc. of the 28th Int. Conf. on Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 1017–1024.
- [81] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, “Deep learning for arabic nlp: A survey,” *Special Issue on The Convergence of New Computing Paradigms and Big Data Analytics Methodologies for Online Social Network*, *J. of Comput. Sci.*, vol. 26, pp. 522–531, May 2018.
- [82] J. Ebrahimi and D. Dou, “Chain based rnn for relation classification,” in *Proc. of the 2015 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, Denver, CO, USA, May 2015, pp. 1244–1249.

- [83] S. Venugopalan, L. A. Hendricks, R. J. Mooney, and K. Saenko, “Improving lstm-based video description with linguistic knowledge mined from text,” 2016, *arXiv:1604.01729*.
- [84] D. Zhang and D. Wang, “Relation classification via recurrent neural network,” 2015, *arXiv:1508.01006*.
- [85] R. Sproat and N. Jaitly, “RNN approaches to text normalization: A challenge,” 2016, *arXiv:1611.00068*.
- [86] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
- [87] M. Sundermeyer, R. Schlüter, and H. Ney, “LSTM neural networks for language modeling,” in *13th Annu. Conf. of the Int. Speech Commun. Assoc.*, Portland, OR, USA, Sep. 2012, pp. 194–197.
- [88] C.-Y. Lin, “Looking for a few good metrics: ROUGE and its evaluation,” in *Proc. of the 4th NTCIR Workshops*, Tokyo, Japan, Jun. 2004, pp. 1–8.
- [89] M. El-Haj, “Arabic multi-document text summarisation,” Ph.D. dissertation, Univ. of Essex, Colchester, UK, 2012.
- [90] C.-Y. Lin and F. J. Och, “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics,” in *Proc. of the 42nd Annu. Meeting of the Assoc. for Comput. Linguistics*, Barcelona, Spain, Jul. 2004, pp. 605–612.
- [91] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. of the 40th Annu. Meeting of the Assoc. for Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.
- [92] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Trans. on Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
- [93] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
- [94] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, “Face recognition: A convolutional neural-network approach,” *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
- [95] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” 2014, *arXiv:1404.2188*.
- [96] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, *arXiv:1408.5882*.

- [97] A. Vedaldi and K. Lenc, “Matconvnet: Convolutional neural networks for matlab,” in *Proc. of the 23rd ACM Int. Conf. on Multimedia*. Brisbane, Australia: ACM, Oct. 2015, pp. 689–692.
- [98] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1725–1732.
- [99] B. Hu, Z. Lu, H. Li, and Q. Chen, “Convolutional neural network architectures for matching natural language sentences,” in *Adv. in Neural Inf. Process. Syst.*, Montreal, Canada, Dec. 2014, pp. 2042–2050.
- [100] K. A. Islam, D. Pérez, V. Hill, B. Schaeffer, R. Zimmerman, and J. Li, “Seagrass detection in coastal water through deep capsule networks,” in *Chin. Conf. on Pattern Recognit. and Comput. Vis.*, Guangzhou, China, Nov. 2018, pp. 320–331.
- [101] D. Pérez, K. Islam, V. Hill, R. Zimmerman, B. Schaeffer, and J. Li, “Deepcoast: Quantifying seagrass distribution in coastal water through deep capsule networks,” in *Chin. Conf. on Pattern Recognit. and Comput. Vis.*, Guangzhou, China, Nov. 2018, pp. 404–416.
- [102] P. Afshar, A. Mohammadi, and K. N. Plataniotis, “Brain tumor type classification via capsule networks,” 2018, *arXiv:1802.10200*.
- [103] A. Jiménez-Sánchez, S. Albarqouni, and D. Mateus, “Capsule networks against medical imaging data challenges,” in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. M.-H. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, and P. Jannin, Eds. Cham, Switzerland: Springer, 2018, pp. 150–160.
- [104] J. Piskorski and R. Yangarber, “Information extraction: Past, present and future,” in *Multi-source, Multilingual Information Extraction and Summarization*, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, Eds. Berlin, Germany: Springer, 2013, pp. 23–49.
- [105] R. Grishman, “Information extraction: Capabilities and challenges,” in *Notes Prepared for the 2012 Int. Winter School in Lang. and Speech Technol.*, Tarragona, Spain, Jan. 2012. Available: <https://cs.nyu.edu/~grishman/tarragona.pdf> Accessed: July 30, 2021
- [106] X. Han and J. Wang, “Earthquake information extraction and comparison from different sources based on web text,” *ISPRS Int. J. of Geo-Inf.*, vol. 8, no. 6, pp. 252–268, May 2019.
- [107] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, “An overview of event extraction from text,” in *Workshop on Detection, Representation, and Exploitation of*

- Events in the Semantic Web (DeRiVE 2011) at 10th Int. Semantic Web Conf.*, Bonn, Germany, Oct. 2011, pp. 48–57.
- [108] H. Wang, “Knowledge base construction from scientific literature,” Ph.D. dissertation, Rensselaer Polytech. Inst., Troy, NY, USA, 2016.
- [109] Y. Liu, T. Zhang, Z. Liang, H. Ji, and D. L. McGuinness, “Seq2rdf: An end-to-end application for deriving triples from natural language text,” 2018, *arXiv:1807.01763*.
- [110] J. Howe, “The rise of crowdsourcing,” *Wired Mag.*, vol. 14, no. 6, pp. 1–4, Jan. 2006.
- [111] M.-C. Yuen, I. King, and K.-S. Leung, “A survey of crowdsourcing systems,” in *2011 IEEE 3rd Int. Conf. on Privacy, Secur., Risk and Trust and 2011 IEEE Third Int. Conf. on Social Comput.*, Boston, MA, USA, Oct. 2011, pp. 766–773.
- [112] M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti, “Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora,” in *Proc. of the Conf. on Empirical Methods in Natural Lang. Process.*, Edinburgh, UK, Jul. 2011, pp. 670–679.
- [113] C. Callison-Burch and M. Dredze, “Creating speech and language data with amazon’s mechanical turk,” in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Lang. Data with Amazon’s Mechanical Turk*, ser. CSLDAMT ’10, Los Angeles, CA, USA, Jun. 2010, pp. 1–12.
- [114] R. Simpson, K. R. Page, and D. De Roure, “Zooniverse: Observing the world’s largest citizen science platform,” in *Proc. of the 23rd Int. Conf. on World Wide Web*, Seoul, Korea, Apr. 2014, pp. 1049–1054.
- [115] M. Buhrmester, T. Kwang, and S. D. Gosling, “Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data?” *Perspectives on Psychological Sci.*, vol. 6, no. 1, pp. 3–5, Feb. 2011.
- [116] M. Keating and R. D. Furberg, “A methodological framework for crowdsourcing in research,” in *Federal Committee on Statist. Methodology Res. Conf.*, Washington, DC, USA, Nov. 2013, pp. 1–8.
- [117] J. M. Leimeister, M. Huber, U. Bretschneider, and H. Krcmar, “Leveraging crowdsourcing: activation-supporting components for it-based ideas competition,” *J. of Manag. Inf. Syst.*, vol. 26, no. 1, pp. 197–224, Summer 2009.
- [118] J. Simpson, E. Weiner, and Oxford University Press, *Oxford English Dictionary*. Oxford, UK: Clarendon Press, 1989.
- [119] E. L. Deci and R. M. Ryan, “Self-determination theory: A macrotheory of human motivation, development, and health,” *Can. Psychol.*, vol. 49, no. 3, pp. 182–185, May 2008.

- [120] W. Mason and D. J. Watts, “Financial incentives and the performance of crowds,” in *Proc. of the ACM SIGKDD Workshop on Human Comput.*, Paris, France, Jun. 2009, pp. 77–85.
- [121] L. de Alfaro and M. Shavlovsky, “Crowdsourcing quantitative evaluation: algorithms and empirical results,” UC Santa Cruz, Santa Cruz, CA, USA, Tech. Rep. UCSC-SOE-14-03, 2017.
- [122] F. Ribeiro, D. Florencio, and V. Nascimento, “Crowdsourcing subjective image quality evaluation,” in *2011 18th IEEE Int. Conf. on Image Process.*, Brussels, Belgium, Dec. 2011, pp. 3097–3100.
- [123] P. Korshunov, S. Cai, and T. Ebrahimi, “Crowdsourcing approach for evaluation of privacy filters in video surveillance,” in *Proc. of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, Nara, Japan, Oct. 2012, pp. 35–40.
- [124] J. M. Mortensen, “Crowdsourcing ontology verification,” in *Int. Semantic Web Conf.*, Sydney, Australia, Oct. 2013, pp. 448–455.
- [125] Y. Zhao and Q. Zhu, “Evaluation on crowdsourcing research: Current status and future direction,” *Inf. Syst. Frontiers*, vol. 16, no. 3, pp. 417–434, Jul. 2014.
- [126] A. Keen, *The Cult of the Amateur*. New York, NY, USA: Currency, 2007.
- [127] J. McCarthy, “Programs with common sense,” in *Symp. on Mechanization of Thought Processes*, Teddington, UK, Nov. 1958, pp. 3–10.
- [128] T. Hofffeld, M. Hirth, and P. Tran-Gia, “Modeling of crowdsourcing platforms and granularity of work organization in future internet,” in *23rd Int. Teletraffic Congr.*, San Francisco, CA, USA, Oct. 2011, pp. 142–149.
- [129] D. C. Brabham, “Crowdsourcing the public participation process for planning projects,” *Planning Theor.*, vol. 8, no. 3, pp. 242–262, Jul. 2009.
- [130] H. Gao, G. Barbier, and R. Goolsby, “Harnessing the crowdsourcing power of social media for disaster relief,” *IEEE Intell. Syst.*, vol. 26, no. 3, pp. 10–14, Jun. 2011.
- [131] J. W. Crampton, “Cartography: Maps 2.0,” *Prog. in Hum. Geog.*, vol. 33, no. 1, pp. 91–100, Feb. 2009.
- [132] A. Hudson-Smith, M. Batty, A. Crooks, and R. Milton, “Mapping for the masses,” *Social Sci. Comput. Rev.*, vol. 27, no. 4, pp. 524–538, Apr. 2009.
- [133] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, “Annotating named entities in twitter data with crowdsourcing,” in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Lang. Data with Amazon’s Mechanical Turk*, Los Angeles, CA, USA, Jun. 2010, pp. 80–88.

- [134] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: A research frontier," *Int. J. of Digit. Earth*, vol. 3, no. 3, pp. 231–241, Sep. 2010.
- [135] Y. Tong, L. Chen, and C. Shahabi, "Spatial crowdsourcing," *Proc. of the VLDB Endowment*, vol. 10, no. 12, pp. 1988–1991, Aug. 2017.
- [136] K. Wazny, "Applications of crowdsourcing in health: An overview," *J. of Global Health*, vol. 8, no. 1, pp. 1–20, Mar. 2018.
- [137] A. Holzinger, "Interactive machine learning for health informatics: When do we need the human-in-the-loop?" *Brain Inform.*, vol. 3, no. 2, pp. 119–131, Mar. 2016.
- [138] A. Filippova, C. Gilroy, R. Kashyap, A. Kirchner, A. C. Morgan, K. Polimis, A. Usmani, and T. Wang, "Humans in the loop: Incorporating expert and crowd-sourced knowledge for predictions using survey data," *Socius: Sociol. Res. for a Dynamic World*, vol. 5, pp. 1–15, Sep. 2019.
- [139] P. Fraternali, A. Castelletti, R. Soncini-Sessa, C. Vaca Ruiz, and A. Rizzoli, "Putting humans in the loop: Social computing for water resources management," *Environ. Model. & Softw.*, vol. 37, pp. 68–77, Nov. 2012.
- [140] I. Ryazanov, A. T. Nylund, D. Basu, I.-M. Hassellv, and A. Schliep, "Deep learning for deep waters: An expert-in-the-loop machine learning framework for marine sciences," *J. of Marine Sci. and Eng.*, vol. 9, no. 2, pp. 1–18, Feb. 2021.
- [141] P. Ristoski, D. Y. Zubarev, A. L. Gentile, N. Park, D. Sanders, D. Gruhl, L. Kato, and S. Welch, "Expert-in-the-loop AI for polymer discovery," in *Proc. of the 29th ACM Int. Conf. on Inf. & Knowl. Manag.*, Virtual Event, Ireland, Oct. 2020, pp. 2701–2708.
- [142] X. Guo, Q. Yu, R. Li, C. O. Alm, C. Calvelli, P. Shi, and A. Haake, "An expert-in-the-loop paradigm for learning medical image grouping," in *Adv. in Knowl. Discovery and Data Mining*, J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, and R. Wang, Eds., Auckland, New Zealand, Apr. 2016, pp. 477–488.
- [143] V. Gerla, V. Kremen, M. Macas, D. Dudysova, A. Mladek, P. Sos, and L. Lhotska, "Iterative expert-in-the-loop classification of sleep PSG recordings using a hierarchical clustering," *J. of Neuroscience Methods*, vol. 317, pp. 61–70, Apr. 2019.
- [144] V. Gerla, V. Kremen, M. Macas, E. Saifutdinova, A. Mladek, and L. Lhotska, "Expert-in-the-loop learning for sleep EEG data," in *2018 IEEE Int. Conf. on Bioinf. and Biomed.*, Madrid, Spain, Dec. 2018, pp. 2590–2596.
- [145] B. D. Weinstein, "What is an expert?" *Theor. Med.*, vol. 14, no. 1, pp. 57–73, Mar. 1993.
- [146] M. Hossain and I. Kauranen, "Crowdsourcing: A comprehensive literature review," *Strategic Outsourcing: An Int. J.*, vol. 8, no. 1, pp. 2–22, Feb. 2015.

- [147] M. K. Poetz and M. Schreier, “The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?” *J. of Product Innov. Manag.*, vol. 29, no. 2, pp. 245–256, Mar. 2012.
- [148] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *Proc. of the 26th Annu. CHI Conf. on Human factors in Comput. Syst.*, Florence, Italy, Apr. 2008, pp. 453–456.
- [149] S. A. Adams, “Sourcing the crowd for health services improvement: The reflexive patient and share-your-experience websites,” *Social Sci. & Med.*, vol. 72, no. 7, pp. 1069–1076, Apr. 2011.
- [150] G. Bugs, C. Granell, O. Fonts, J. Huerta, and M. Painho, “An assessment of public participation gis and web 2.0 technologies in urban planning practice in Canela, Brazil,” *Cities*, vol. 27, no. 3, pp. 172–181, Jun. 2010.
- [151] A. Wiggins and K. Crowston, “From conservation to crowdsourcing: A typology of citizen science,” in *2011 44th Hawaii Int. Conf. on Syst. Sci.*, Kauai, HI, USA, Jan. 2011, pp. 1–10.
- [152] K. Muthukumaraswamy, “When the media meet crowds of wisdom,” *Journalism Pract.*, vol. 4, no. 1, pp. 48–65, Feb. 2010.
- [153] D. Hummer, “The carbon mineral challenge: A worldwide effort to find earths missing carbon minerals,” *Australian J. of Mineralogy*, vol. 20, no. 1, pp. 55–63, Nov. 2019.
- [154] M. Maskey, R. Ramachandran, and J. Miller, “Deep learning for phenomena-based classification of earth science images,” *J. of Applied Remote Sensing*, vol. 11, no. 4, p. 1, Sep. 2017.
- [155] Z. Zhai, T. Kijewski-Correa, D. Hachen, and G. Madey, “Haiti earthquake photo tagging: Lessons on crowdsourcing in-depth image classifications,” in *7th Int. Conf. on Digit. Inf. Manag.*, Macau, Macao, Aug. 2012, pp. 357–364.
- [156] S. C. D’Souza, “Parser extraction of triples in unstructured text,” 2018, *arXiv:1811.05768*.
- [157] N. Kertkeidkachorn and R. Ichise, “T2KG: An end-to-end system for creating knowledge graph from unstructured text,” in *Workshops at the 31st AAAI Conf. on Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 743–749.
- [158] R. Rastogi, “Building knowledge bases from the web,” in *Proc. of the 18th Int. Conf. on Manag. of Data*, Pune, India, Dec. 2012, pp. 5–9.
- [159] P. Exner and P. Nugues, “Entity extraction: From unstructured text to DBpedia RDF triples,” in *Proc. of the Web of Linked Entities Workshop in Conjunction with the 11th Int. Semantic Web Conf.*, Boston, MA, USA, Nov. 2012, pp. 58–69.

- [160] R. J. Mooney and R. Bunescu, “Mining knowledge from text using information extraction,” *ACM SIGKDD Explor. Newslett.*, vol. 7, no. 1, pp. 3–10, Jun. 2005.
- [161] A. Voutilainen, “Part-of-speech tagging,” in *The Oxford Handbook of Computational Linguistics*, R. Mitkov, Ed. Oxford, UK: Oxford Univ. Press, Sep. 2003, pp. 219–232.
- [162] J. Nivre, “Dependency grammar and dependency parsing,” Växjö University, School of Mathematics and Systems Engineering, Tech. Rep. MSI 05133, 2005. Available: <http://stp.lingfil.uu.se/~nivre/docs/05133.pdf> Accessed: July 29, 2021
- [163] K. Clark and C. D. Manning, “Deep reinforcement learning for mention-ranking coreference models,” in *Proc. of the 2016 Conf. on Empirical Methods on Natural Lang. Process.*, Austin, TX, USA, Nov. 2016, pp. 2256–2262.
- [164] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo, 2020. Available: <https://zenodo.org/record/3358113> Accessed: July 31, 2021
- [165] L. Del Corro and R. Gemulla, “Clausie: clause-based open information extraction,” in *Proc. of the 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, May 2013, pp. 355–366.
- [166] E. Chourdakis and J. Reiss, “Grammar informed sound effect retrieval for soundscape generation,” in *DMRN+ 13: Digit. Music Res. Netw. 1-day Workshop*, London, UK, Nov. 2018, p. 9.
- [167] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” 2017, *arXiv:1711.00043*.
- [168] T. Deselaers, S. Hasan, O. Bender, and H. Ney, “A deep learning approach to machine transliteration,” in *Proc. of the 4th Workshop on Statist. Mach. Transl.*, Athens, Greece, Mar. 2009, pp. 233–241.
- [169] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar *et al.*, “Tensor2tensor for neural machine translation,” 2018, *arXiv:1803.07416*.
- [170] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain, “Machine translation using deep learning: An overview,” in *2017 Int. Conf. on Comput., Commun. and Electron.*, Jaipur, India, Jul. 2017, pp. 162–167.
- [171] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, *arXiv:1409.0473*.
- [172] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, “Deep learning for answer sentence selection,” 2014, *arXiv:1412.1632*.
- [173] Y. Kim, Y. Jernite, D. Sontag, and A. Rush, “Character-aware neural language models,” in *AAAI Conf. on Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 2741–2749.

- [174] C. D. Santos and B. Zadrozny, “Learning character-level representations for part-of-speech tagging,” in *Proc. of the 31st Int. Conf. on Mach. Learn.*, Beijing, China, Jun. 2014, pp. 1818–1826.
- [175] F. Peng, D. Schuurmans, S. Wang, and V. Keselj, “Language independent authorship attribution using character level language models,” in *Proc. of the 10th Conf. on Eur. Chapter of the Assoc. for Comput. Linguistics*, ser. EACL ’03, Budapest, Hungary, Apr. 2003, pp. 267–274.
- [176] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, “Parsing natural scenes and natural language with recursive neural networks,” in *Proc. of the 28th Int. Conf. on Mach. Learn.*, Bellevue, WA, USA, Jun. 2011, pp. 129–136.
- [177] A. Severyn and A. Moschitti, “Twitter sentiment analysis with deep convolutional neural networks,” in *Proc. of the 38th Int. ACM SIGIR Conf. on Res. and Develop. in Inf. Retrieval*, Santiago, Chile, Aug. 2015, pp. 959–962.
- [178] H. Zhao, Z. Lu, and P. Poupard, “Self-adaptive hierarchical sentence model,” in *24th Int. Joint Conf. on Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 4069–4076.
- [179] X. Han, S. Cao, L. Xin, Y. Lin, Z. Liu, M. Sun, and J. Li, “OpenKE: An open toolkit for knowledge embedding,” in *Proc. of EMNLP*, Brussels, Belgium, Nov. 2018, pp. 139–144.
- [180] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” 2014, *arXiv:1409.3215*.
- [181] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” 2016, *arXiv:1609.08144*.
- [182] S. Banerjee and A. Lavie, “METEOR: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Eval. Measures for Mach. Transl. and/or Summarization*, Ann Arbor, MI, USA, Jun. 2005, pp. 65–72.
- [183] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, “Video description: A survey of methods, datasets, and evaluation metrics,” *ACM Comput. Surveys*, vol. 52, no. 6, pp. 1–37, Jan. 2020.
- [184] J. Brank, M. Grobelnik, and D. Mladenic, “A survey of ontology evaluation techniques,” in *Proc. of the Conf. on Data Mining and Data Warehouses*, Ljubljana, Slovenia, 2005, pp. 166–170.
- [185] D. Vrandečić, “Ontology evaluation,” in *Handbook on Ontologies*, S. Staab and R. Studer, Eds. Berlin, Heidelberg: Springer, 2009, pp. 293–313.

- [186] A. Maedche and S. Staab, “Measuring similarity between ontologies,” in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. Berlin, Germany: Springer, 2002, pp. 251–263.
- [187] R. Porzel and R. Malaka, “A task-based approach for ontology evaluation,” in *ECAI Workshop on Ontology Learn. and Population*, Valencia, Spain, Aug. 2004, pp. 1–6.
- [188] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, “Data driven ontology evaluation,” in *Proc. of the 4th Int. Conf. on Lang. Resour. and Eval.*, Lisbon, Portugal, May 2004, pp. 641–644.
- [189] A. Lozano-Tello and A. Gomez-Perez, “Ontometric,” *J. of Database Manag.*, vol. 15, no. 2, pp. 1–18, Apr 2004.
- [190] A. Burton-Jones, V. C. Storey, V. Sugumaran, and P. Ahluwalia, “A semiotic metrics suite for assessing the quality of ontologies,” *Data & Knowl. Eng.*, vol. 55, no. 1, pp. 84 – 102, Oct. 2005.
- [191] P. Grogono, A. Batarekh, A. Preece, R. Shinghal, and C. Suen, “Expert system evaluation techniques: A selected bibliography,” *Expert Syst.*, vol. 8, no. 4, pp. 227–239, Nov. 1991.
- [192] G. A. Miller, “Wordnet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.
- [193] A. Rajaraman and J. D. Ullman, “Data mining,” in *Mining of Massive Datasets*. Cambridge, UK: Cambridge Univ. Press, 2011, pp. 1–17.
- [194] S. Gehrmann, Y. Deng, and A. M. Rush, “Bottom-up abstractive summarization,” 2018, *arXiv:1808.10792*.
- [195] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” 2017, *arXiv:1705.04304*.
- [196] R. Nallapati, F. Zhai, and B. Zhou, “SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents,” in *31st AAAI Conf. on Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 3075–3081.
- [197] I. Androutsopoulos and P. Malakasiotis, “A survey of paraphrasing and textual entailment methods,” *J. of Artif. Intell. Res.*, vol. 38, no. 1, pp. 135–187, May 2010.
- [198] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, “AllenNLP: A deep semantic natural language processing platform,” 2017, *arXiv:1803.07640*.

APPENDIX A

IRB APPROVAL FOR CROWDSOURCING EXPERIMENT

Below you will see both the initially IRB proposal for the crowdsourcing experiment, and the renewal of the IRB proposal submitted after the time limit of the initial approval was done.

Consent to Participate in Research

You are invited to participate in a Research study that has been approved by the Rensselaer Institutional Review Board (IRB). The IRB reviews and approves all human subject research in accordance with applicable state law and federal law governing Human Subject Research.

RESEARCH STUDY TITLE: WRITING RULES FOR A SELECTED NARRATIVE(S).

PRINCIPAL INVESTIGATOR: ANIRUDH PRABHU

DESCRIPTION

The study will present you with a story and ask you to write rules about the world described in the story.

METHODS

- Data is collected through a through a survey-like format. You will first be randomly assigned a story. After reading the story, the system will present you with multiple text boxes to write rules that important to the story, according to you. You are required to enter a minimum of 7 rules for each story. The estimated duration of the task will be between 30-75 mins depending on the length of the story and your reading speed.

BENEFITS

There is no direct benefit to you outside of the compensation gained. But the knowledge gained from the study will help us explore how individuals understand a story.

POSSIBLE RISKS AND DISCOMFORT

- There are no expected risks or discomfort involved in this study.

PAYMENT AND COMPENSATION

- In return for your participation, you will receive up to \$0.50 through the mechanical turk platform.
- The exact amount for the compensation will be decided by the length of the story assigned to you.

PARTICIPANT'S RIGHTS

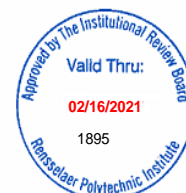
- Your participation in this research study is voluntary. You may end your participation at any time without penalty.

DATA CONFIDENTIALITY AND SECURITY

- No information regarding your name or other personally identifiable information will be collected during the survey. All data collected will be stored in a password protected computer in a locked room, accessible to the researchers involved in the study. All data will be retained for a minimum period of three years after the completion of data collection. Under no circumstances will the personally identifiable information be disclosed.

SHARING OF RESEARCH RESULTS

- The results of this study may be presented at scientific or professional meetings, published in scientific journals, or otherwise shared with the general public. However, no statements will



be directly attributed to you, and your name or other identifying information will not be publicly shared or included in any published materials.

CONTACT INFORMATION

For further information or questions about this research, please contact:

Mr. Anirudh Prabhu, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, Email: prabha2@rpi.edu

Dr. Peter Fox, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, Email: pfox@cs.rpi.edu

Chair, Institutional Review Board, Rensselaer Polytechnic Institute, CII 9015, 110 8th Street, Troy, NY 12180 Phone: 518- 276-4873, Email: irb@rpi.edu

SIGNATURE 1: CONSENT TO PARTICIPATE IN RESEARCH

By signing below, I confirm that I am at least 18 years of age and that I consent to participation in the research study described above.

Signed _____ Date _____

Printed Name _____

Study Participants: Please keep the second copy of this Consent Form for your records.

1/30/2020

Page 2 of 3

Note: Do not sign this consent form if it does not have the IRB approval stamp, or if the date has lapsed.



Consent to Participate in Research

You are invited to participate in a Research study that has been approved by the Rensselaer Institutional Review Board (IRB). The IRB reviews and approves all human subject research in accordance with applicable state law and federal law governing Human Subject Research.

RESEARCH STUDY TITLE: WRITING RULES FOR A SELECTED NARRATIVE(S).

PRINCIPAL INVESTIGATOR: ANIRUDH PRABHU

DESCRIPTION

The study will present you with a story and ask you to write rules about the world described in the story.

METHODS

- Data is collected through a through a survey-like format. You will first be randomly assigned a story. After reading the story, the system will present you with multiple text boxes to write rules that important to the story, according to you. You are required to enter a minimum of 7 rules for each story. The estimated duration of the task will be between 30-75 mins depending on the length of the story and your reading speed.

BENEFITS

There is no direct benefit to you outside of the compensation gained. But the knowledge gained from the study will help us explore how individuals understand a story.

POSSIBLE RISKS AND DISCOMFORT

- There are no expected risks or discomfort involved in this study.

PAYMENT AND COMPENSATION

- In return for your participation, you will receive up to \$0.50 through the mechanical turk platform.
- The exact amount for the compensation will be decided by the length of the story assigned to you.

PARTICIPANT'S RIGHTS

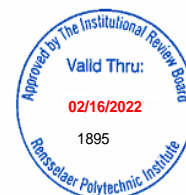
- Your participation in this research study is voluntary. You may end your participation at any time without penalty.

DATA CONFIDENTIALITY AND SECURITY

- No information regarding your name or other personally identifiable information will be collected during the survey. All data collected will be stored in a password protected computer in a locked room, accessible to the researchers involved in the study. All data will be retained for a minimum period of three years after the completion of data collection. Under no circumstances will the personally identifiable information be disclosed.

SHARING OF RESEARCH RESULTS

- The results of this study may be presented at scientific or professional meetings, published in scientific journals, or otherwise shared with the general public. However, no statements will



be directly attributed to you, and your name or other identifying information will not be publicly shared or included in any published materials.

CONTACT INFORMATION

For further information or questions about this research, please contact:

Mr. Anirudh Prabhu, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, Email: prabha2@rpi.edu

Dr. Peter Fox, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, Email: pfox@cs.rpi.edu

Chair, Institutional Review Board, Rensselaer Polytechnic Institute, CII 9015, 110 8th Street, Troy, NY 12180 Phone: 518- 276-4873, Email: irb@rpi.edu

SIGNATURE 1: CONSENT TO PARTICIPATE IN RESEARCH

By signing below, I confirm that I am at least 18 years of age and that I consent to participation in the research study described above.

Signed _____ Date _____

Printed Name _____

Study Participants: Please keep the second copy of this Consent Form for your records.

1/30/2020

Page 2 of 3

Note: Do not sign this consent form if it does not have the IRB approval stamp, or if the date has lapsed.



APPENDIX B

CROWDSOURCING EXPERIMENT IN MECHANICAL TURK

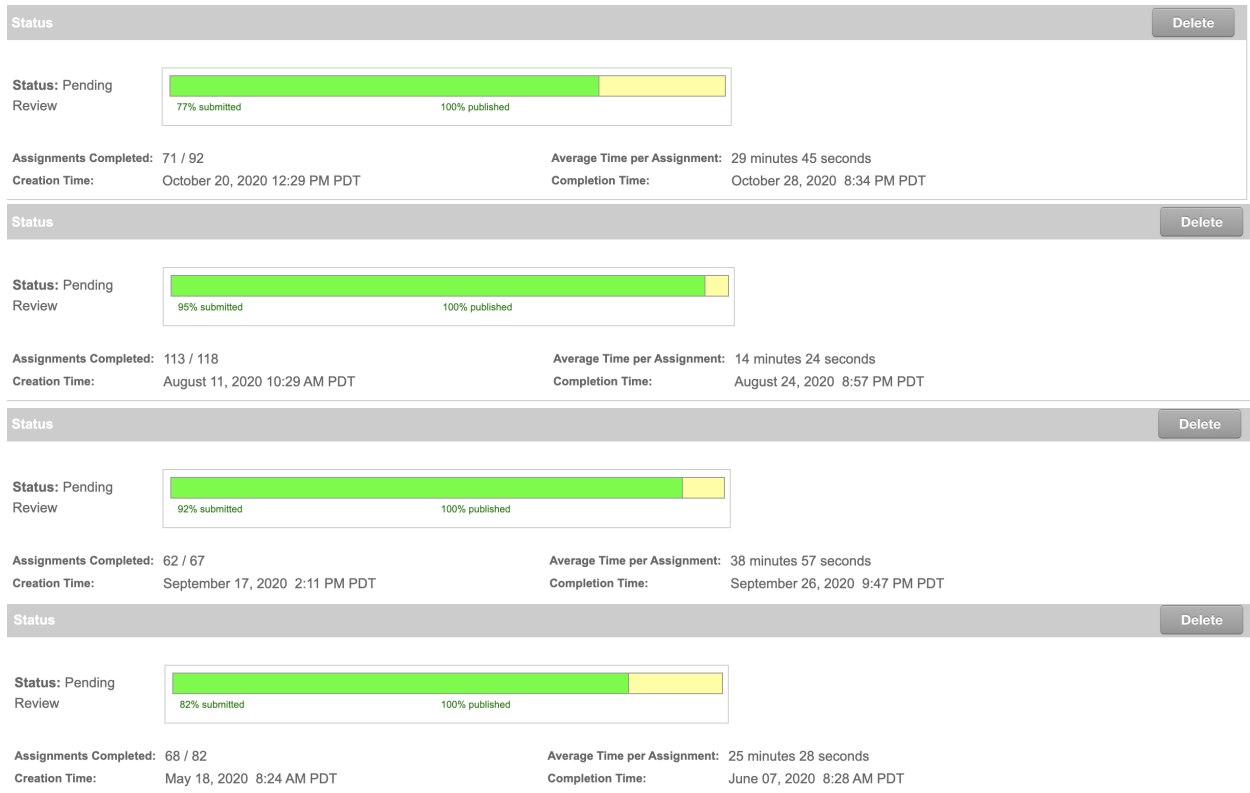


Figure B.1: Examples of the batches published in Amazon Mechanical Turk. Each batch summary shows the average time taken per task. Number of tasks (assignments), the creation and completion time of each task.

Customize View Filter Results Upload CSV Approve All Download CSV

Approve Reject

Input.Full Url	Input.Final Bonus	Rules1	Rules10	Rules11	Rules12	Rules13	Rules14	Rules15	Rules16	Rules17	Rules18	R
http://orion.tw.rpi.edu/~anirudhprabhu/ShortSto...	0.45	The manner in which a man has lived is often th...										
http://orion.tw.rpi.edu/~anirudhprabhu/ShortSto...	0.45	If Donegal pretended that he didn't know about ...										
http://orion.tw.rpi.edu/~anirudhprabhu/ShortSto...	0.45	Paragraph	Context of the story	Final note of the author	Creativity	Intriguing ending, raising doubts about what el...	Credits to participants	Author's personal style	Verbs and standards			
http://orion.tw.rpi.edu/~anirudhprabhu/ShortSto...	0.45	Donegal must walk with a cane.										
http://orion.tw.rpi.edu/~anirudhprabhu/ShortSto...	0.45	There were two important things-one, that she ...	Peter was most dreadfully frightened; he rushed...	Once upon a time there was a village shop. The ...	"Swearin'! 'E! Why, 'e don't know any."	In the centre of the great city of London lies ...	_MY DEAR FREDA,_ _Because you are fond of ...	[[Illustration] One morning a little rabbit sat...	_Did you ever wonder at the lonely life the bir...	[Transcriber's Note: This etext was produced fr...	_The starways' Lone Watcher had expected some o...	Tin Trotz only to ...

Figure B.2: Examples of the results in Amazon Mechanical Turk. The figure shows the interface presented to requester for approving or rejecting rules.

Instructions

Summary	Detailed Instructions	Examples
----------------	-----------------------	----------

You will be assigned with a story to read. Once you read the story you must come up with at least 5 rules that you see obeyed in the story.

Instructions

Summary	Detailed Instructions	Examples
---------	------------------------------	----------

More details

Each person is assigned a random story. You must read the story completely. As you read the story, you will start noticing rules that govern the world described in that story. If you identify a rule that affects the result of the story or changes the events described in the story, please note it down and compile in the text boxes provided to you. At least 5 rules must be presented per story. The longer the story, the more potential for finding rules.

In the examples section, we use a commonly children's known of "The Three little pigs and the big bad wolf" to show examples of the rules that can be obtained from them.

Bonuses will be provided for every task. Bonus amounts are decided based on the length of the story assigned, and thus have been precalculated. The longer the story, the higher the bonus, upto a maximum of \$0.45

Instructions

Summary	Detailed Instructions	Examples
---------	-----------------------	-----------------

Good examples	Bad examples
<ul style="list-style-type: none"> • If the wolf sees pigs, then wolf will chase after the pigs. • If pigs are in the house, then wolf cannot eat pigs. • If house is built with heavy materials, then wolf cannot blow it down. • Light houses are built quickly. • Heavy houses are not built quickly. <p>Be precise with your rules.</p>	<ul style="list-style-type: none"> • There were 3 pigs. (A fact, not a rule) • If wolf sees pigs, then wolf will chase and eat it once it is caught. (Complex rule, can be split into 2 rules) • It takes a pig a whole day to build a house with bricks. (A fact, not a rule) • You should not build a straw house. (A conclusion, not a pattern) <p>Do not include more information than required into a single rule.</p> <p>If the rule seems too complicated, then check if it can be split into 2 rules.</p> <p>Avoid writing facts (things that happen in the story), instead read between the lines and find patterns.</p> <p>Do not write conclusions.</p> <p>No typos please, entries with typos will be rejected.</p>

Figure B.4: Instructions from the Amazon Mechanical Turk interface that are presented to the participants of the crowdsourcing experiment.

APPENDIX C

CROWDSOURCING RULES EXAMPLES FROM PUBLISHED BATCHES IN MTURK

Story : A Hitch in Space

- Jeff Bogart can create an identical clone of a person
- At any given time only one of the clones was visible/available for Jeff Bogart.
- The real Joseph and imaginary Joseph never exist in the same room together
- The ship could not exceed a velocity of 15 miles per second
- Joseph had to stay away from the ships jet or else it would destroy him
- Joseph could not stay out in space longer than the suit oxygen that remained

Story : A Witch in Time

- If nat hadn't used her paralysis ray, abby would have been hanged
- If you are accused of witchcraft, you will be hanged.
- if abby hadn't traveled with nat to other times, nat wouldn't have been blamed for Time Meddling
- If you stay too long on the sun, your skin will tan.
- If you have protection, you won't be affected by a paralysis ray.
- If you are hit by the paralysis ray, you'll lose consciousness
- If you have a time machine, you can travel through time
- If you are from the 17th century, you won't be able to speak 25th century language
- If viewers hadn't been frozen like statues, they would have seen abby's body being repositioned to be hanged

Story : A Gift from Earth

- If the Earthmen come, the people of Zur will talk about them.
- If the youngest brother speaks at the conference, the other brothers become annoyed.
- If anyone on Zur acquires metal, then he or she will become rich.
- Anti-factions exist to oppose any suggested changes.
- Humans use the guise of cooperation to take advantage of one another.
- Zurians and Humans are identical physically.
- If you are born in Zur, you are not an Earthman
- If you are an Earthman, you do not speak the language of Zur
- If Earthman find metal in another world, they will take it
- If you are a Earthwoman, you attract Zurian men easily
- If you are an Earthman, you come from an overpopulated planet

Story : Accidental Death

- Keep climbing even if you keep slipping
- To program the ship for a star-jump, you merely told it where you were and where you wanted to go
- To program the ship for a star-jump, you merely told it where you were and where you wanted to go
- Time in space seems shorter than time on earth
- Terminal velocity for a human body falling through air is about one hundred twenty m.p.h.
- The natives of Chang are hostile towards humans.
- If that suit runs out of oxygen, people will lose consciousness due to anoxia.

- If you switch on the emergency tank, it will bring you back.
- If you crack the suit, you can breathe fresh air.

Story : A Pail of Air

- There is no sound in a vacuum.
- the fire must never go out
- Earth loses sun's protection if it moves away from it
- Earth gets very cold if taken away from the sun.
- if someone got no air or heat, they freeze to die
- if the character and her Pa and Ma see someone else, they will get scared
- there is vacuum outside the nest, so the air must be taken by pails
- once there is no sun and no moon, they must use clocks to remind them of time
- Always keep a big reserve of air.
- Suit is required to be worn when going out of the nest.

APPENDIX D

RESULTS FROM CROWDSOURCING EVALUATION

Displayed below are the results from the Crowdsourcing Evaluation from chapter 4. The tables contain the following headers:

Table D.1: Data description for crowdsourcing evaluation.

Title	Description
RuleSet	There are 20 rulesets in each batch sent for evaluation. This parameter keeps track of the rule set. This is used the unique ID for the evaluation table.
Story	This parameter links to the original story for which these rules were generated. The participants do NOT need to read the story to evaluate the rules, these story links are included for easing further processing and linking rule bases to the fact bases.
Rule	The rule output by the mechanical turk participants. This is the result that need to be evaluated.
Accepted/Rejected	If there is context for what concept the pronoun is referring to, then the rule is acceptable. If there is no context for the concept being referred to with the pronouns, the rule should be marked as rejected.
Simple/Complex	A "simple" rule, where a condition/fluent is stated to be true or false for the given world. And a "complex" rule, where a conditional trigger is presented and the result of how the world is affected by that trigger being fulfilled is also document.
CSR/NCSR	The Common Sense Rule (CSR) labeling is done based on whether the rule is understood by the evaluator without having read the story or without being provided any context for the story. If the rule does not follow common sense, or the evaluator needs to actually read the story to verify the validity of the rule, then the rule is marked as Not a Common Sense Rule (NCSR).
Fixed Rule	This column contains the rules that have been corrected by the evaluators during their review.

Table D.2 shows a preview of the crowdsourcing evaluation. The entire dataset can be found at "<https://zenodo.org/record/5154094>".

Table D.2: Data preview for crowdsourcing evaluation.

RuleSet	Story	Rule	Accepted/Rejected	Simple/Complex	CSR/NCSR	Fixed Rule
1	Consignment	If you go in the jungle, the man-killer will find you.	Rejected	Complex	NCSR	Undecided
1	Consignment	The prisoner did a bad thing and now paying for it.	Accepted	Simple	CSR	The prisoner is doing time for his crime.
1	Consignment	If the cops know the prisoner's break out, they will find them.	Accepted	Complex	NCSR	If cops know the prisoner broke out, then they will find him..
1	Consignment	If you run on foot for miles, your feet will hurt.	Accepted	Complex	CSR	If you run for miles, then your feet will hurt.
1	Consignment	Cops are trained to find prisoners who escape	Accepted	Simple	NCSR	Undecided
1	Consignment	The car was going too fast.	Rejected	Simple	NCSR	Undecided (superlative)

APPENDIX E

TEXT TO RULE CONVERSION FULL RESULTS

Table E.1: Data description for text to rules conversion results.

Title	Description
RuleSet	There are 20 rulesets in each batch sent for evaluation. This parameter keeps track of the rule set. This is used the unique ID for the evaluation table.
Story	This parameter links to the original story for which these rules were generated. The participants do NOT need to read the story to evaluate the rules, these story links are included for easing further processing and linking rule bases to the fact bases.
Rule	The rule output by the mechanical turk participants. This is the result that need to be evaluated.
SitCalcRule	This column contains the rules that have been converted to Situation Calculus Rules written in Golog using the Text2Rules method described in chapter 5.

Table E.2: Data preview for text to rules conversion results.

RuleSet	Story	Rule	SitCalcRule
1	A Child’s Dream of a Star	If children looked at the sky, children would always see a bright star.	looked(children, at the sky) :- see(children, a bright star).
1	A Child’s Dream of a Star	Looking at the Bright Star made children imagine many things.	imagine(children, many things).
1	A Child’s Dream of a Star	When the boy looked at the Bright Star the boy remembers the boy sister who died at a young age.	looked(the boy, When) :- died(who, at a young age).
3	A Child’s Dream of a Star	Everything happens at the right time.	happens(Everything, at the right time).
3	A Child’s Dream of a Star	Death is inevitable.	is(Death, inevitable).
3	A Child’s Dream of a Star	Patience is a virtue.	is(Patience, a virtue).

Table E.2 shows a preview of the text to rules conversion results. The entire dataset can be found at ”<https://zenodo.org/record/5154116>”.

APPENDIX F

EVALUATION RESULTS FOR THE SAAG WORKFLOW

F.1 Interpretability, Coverage and Normalized Coverage

Table F.1: Data description for evaluation results for the SAAG workflow.

Title	Description
Story	This parameter links to the name of the original story for which these results were obtained.
Coverage	Coverage measures how well the rule base covers the known laws of the selected world to be modeled.
Coverage normalized	The coverage score can be normalized by dividing the obtained score from the previous step by the count of the facts commonly occurring in the knowledge base and the rule base.
Interpretability	Interpretability measures if the concepts used in the rules are meaningful and/or exist in the knowledge base.

Table F.2: Data preview for evaluation results for the SAAG workflow.

	unique.MBR_SitCalcRules.Story.	Coverage	Coverage_normalized	Interpretability
1	A Child's Dream of a Star	-2.77258872223978	-Inf	0
2	A Cold Night for Crying	-49.9065970003161	-Inf	0.113636363636364
3	A Gift from Earth	-33.9642118474373	-Inf	0.0714285714285714
4	A Hitch in Space	-31.8847703057575	-Inf	0.0634920634920635
5	A Little Journey	-12.476649250079	-Inf	0.0555555555555556
6	A Matter of Order	-21.4875625973583	-Inf	0.121212121212121

The complete results for all the stories used in the SAAG workflow can be found at "<https://zenodo.org/record/5154156>".

F.2 Lawfulness

Attached below is the detailed output of loading all the rule bases generated using the SAAG workflow into SWI-Prolog. There are 204 errors produced, which means that 930 lawful rules generated in 65 rule bases by implementing the SAAG workflow.

```

1 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Child's Dream of
   a Star.pl:5:
2 No permission to modify static procedure '(is)/2'
```

```
3 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Child's Dream of
  a Star.pl:6:
4   No permission to modify static procedure '(is)/2'
5 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Child's Dream of
  a Star.pl:10:
6   No permission to modify static procedure '(is)/2'
7 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Cold Night for
  Crying.pl:1:
8   Arithmetic: 'cold' is not a function
9 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Cold Night for
  Crying.pl:7:3: Syntax error: Operator expected
10 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Cold Night for
  Crying.pl:11:3: Syntax error: Operator expected
11 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Gift from Earth.
  pl:11:
12   No permission to modify static procedure '(is)/2'
13 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Gift from Earth.
  pl:12:
14   No permission to modify static procedure '(is)/2'
15 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Gift from Earth.
  pl:13:
16   No permission to modify static procedure '(is)/2'
17 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :11:3: Syntax error: Operator expected
18 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :13:
19   No permission to modify static procedure '(is)/2'
20 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :16:
21   No permission to modify static procedure '(is)/2'
22 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :17:
23   Arithmetic: 'strange' is not a function
24 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :18:
25   No permission to modify static procedure '(is)/2'
26 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Hitch in Space.pl
  :20:
27   No permission to modify static procedure '(is)/2'
```

```

28 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Little Journey.pl
    :6:
29   Arithmetic: 'expensive' is not a function
30 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Little Journey.pl
    :12:17: Syntax error: Operator expected
31 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Pail of Air.pl:2:
32   No permission to modify static procedure '(is)/2'
33 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Pail of Air.pl:9:
34   No permission to modify static procedure '(is)/2'
35 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Pail of Air.pl
    :11:3: Syntax error: Operator expected
36 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Pail of Air.pl
    :13:3: Syntax error: Operator expected
37 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:1:
38   Arithmetic: 'yourpresentaddress' is not a function
39 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:3:
40   No permission to modify static procedure '(is)/2'
41 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:5:
42   No permission to modify static procedure '(is)/2'
43 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:6:
44   No permission to modify static procedure '(is)/2'
45 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:7:
46   No permission to modify static procedure '(is)/2'
47 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:10:
48   No permission to modify static procedure '(is)/2'
49 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Traveler in Time.
    pl:12:1: Syntax error: End of file in quoted atom
50 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/A Witch in Time.pl
    :4:1: Syntax error: End of file in quoted atom
51 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental Death.pl
    :1:3: Syntax error: Operator expected
52 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental Death.pl
    :4:4: Syntax error: Operator expected

```

```

53 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental_Death.pl
    :5:3: Syntax error: Operator expected
54 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental_Death.pl
    :9:
55   No permission to modify static procedure '(is)/2'
56 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental_Death.pl
    :10:
57   No permission to modify static procedure '(is)/2'
58 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Accidental_Death.pl
    :14:1: Syntax error: End of file in quoted atom
59 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/All_Cats_Are_Gray.
    pl:2:1: Syntax error: End of file in quoted atom
60 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/All_Jackson's
    Children.pl:6:1: Syntax error: End of file in quoted atom
61 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/All_The_People.pl
    :3:
62   No permission to modify static procedure '(is)/2'
63 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/All_The_People.pl
    :7:1: Syntax error: End of file in quoted atom
64 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Aloys.pl:1:3:
    Syntax error: Operator expected
65 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Aloys.pl:7:3:
    Syntax error: Operator expected
66 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Aloys.pl:13:
67   No permission to modify static procedure '(is)/2'
68 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Aloys.pl:15:
69   Arithmetic: 'veryintelligent' is not a function
70 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Am_I_Still_There?.
    pl:2:
71   No permission to modify static procedure '(is)/2'
72 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Am_I_Still_There?.
    pl:6:
73   No permission to modify static procedure '(is)/2'
74 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Am_I_Still_There?.
    pl:15:
75   Arithmetic: 'experimental' is not a function
76 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An_Incident_on
    Route_12.pl:4:3: Syntax error: Operator expected

```

```
77 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Incident on
    Route 12.pl:8:1: Syntax error: End of file in quoted atom
78 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:5:
79 No permission to modify static procedure '(is)/2'
80 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:14:3: Syntax error: Operator expected
81 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:19:3: Syntax error: Operator expected
82 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:24:
83 Arithmetic: 'received' is not a function
84 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:25:
85 No permission to modify static procedure '(is)/2'
86 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:27:36: Syntax error: illegal_character
87 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:29:3: Syntax error: Operator expected
88 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Occurrence at
    Owl Creek.pl:31:1: Syntax error: End of file in quoted atom
89 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Ounce of Cure.pl
    :2:34: Syntax error: Operator expected
90 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Ounce of Cure.pl
    :9:
91 No permission to modify static procedure '(is)/2'
92 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/An Ounce of Cure.pl
    :12:1: Syntax error: End of file in quoted atom
93 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/And All the Earth a
    Grave.pl:4:25: Syntax error: Operator expected
94 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/And All the Earth a
    Grave.pl:8:3: Syntax error: Operator expected
95 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/And All the Earth a
    Grave.pl:15:2: Syntax error: Unexpected end of file
96 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Arm of the Law.pl
    :10:
97 No permission to modify static procedure '(is)/2'
98 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Arm of the Law.pl
    :11:
```

```

99  No permission to modify static procedure '(is)/2'
100 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Arm of the Law.pl
    :14:
101  No permission to modify static procedure '(is)/2'
102 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/As Long As You Wish
    .pl:4:
103  Arithmetic: 'ofsuperiormakeandsyntheticsource' is not a function
104 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/As Long As You Wish
    .pl:5:
105  Arithmetic: 'what' is not a function
106 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/As Long As You Wish
    .pl:6:
107  No permission to modify static procedure '(is)/2'
108 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/As Long As You Wish
    .pl:8:
109  No permission to modify static procedure '(is)/2'
110 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Ask a Foolish
    Question.pl:4:
111  No permission to modify static procedure '(is)/2'
112 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Ask a Foolish
    Question.pl:15:3: Syntax error: Unexpected end of file
113 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/B-12's Moon Glow.pl
    :12:
114  No permission to modify static procedure '(is)/2'
115 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/B-12's Moon Glow.pl
    :28:
116  No permission to modify static procedure '(is)/2'
117 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Medicine.pl:7:
118  No permission to modify static procedure '(is)/2'
119 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Medicine.pl:9:
120  No permission to modify static procedure '(is)/2'
121 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Medicine.pl:11:
122  No permission to modify static procedure 'call/2'
123  Defined at /Applications/SWI-Prolog.app/Contents/swipl/boot/init.pl:310
124 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Medicine.pl
    :14:1: Syntax error: End of file in quoted atom
125 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Memory.pl:10:
126  Arithmetic: 'greater' is not a function
127 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bad Memory.pl:14:

```

```

128 No permission to modify static procedure '(is)/2'
129 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Belly Laugh.pl:1:
130 No permission to modify static procedure '(is)/2'
131 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Belly Laugh.pl:8:3:
    Syntax error: Operator expected
132 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Belly Laugh.pl:15:
133 No permission to modify static procedure '(is)/2'
134 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Belly Laugh.pl
    :21:3: Syntax error: Operator expected
135 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beside Still Waters
    .pl:10:3: Syntax error: Operator expected
136 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beside Still Waters
    .pl:12:
137 No permission to modify static procedure '(is)/2'
138 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beside Still Waters
    .pl:13:
139 No permission to modify static procedure '(is)/2'
140 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Lies the Wub
    .pl:4:
141 No permission to modify static procedure '(is)/2'
142 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Lies the Wub
    .pl:8:3: Syntax error: Operator expected
143 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Lies the Wub
    .pl:10:
144 No permission to modify static procedure '(is)/2'
145 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Lies the Wub
    .pl:11:3: Syntax error: Operator expected
146 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Lies the Wub
    .pl:17:3: Syntax error: Operator expected
147 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond Pandora.pl
    :4:1: Syntax error: End of file in quoted atom
148 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Beyond the Door.pl
    :18:
149 Arithmetic: 'there' is not a function
150 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Black Eyes and the
    Daily Grind.pl:1:
151 No permission to modify static procedure '(is)/2'
152 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Black Eyes and the
    Daily Grind.pl:16:1: Syntax error: End of file in quoted atom

```

```
153 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bread_Overhead.pl
      :1:12: Syntax error: Operator expected
154 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bread_Overhead.pl
      :7:52: Syntax error: Operator expected
155 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bread_Overhead.pl
      :11:
156   Arithmetic: 'relevant' is not a function
157 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bread_Overhead.pl
      :18:
158   No permission to modify static procedure 'float/1'
159 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breakdown.pl:8:3:
      Syntax error: Operator expected
160 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breakdown.pl:10:3:
      Syntax error: Operator expected
161 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breakdown.pl:12:1:
      Syntax error: End of file in quoted atom
162 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :1:
163   Arithmetic: 'areason' is not a function
164 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :2:3: Syntax error: Operator expected
165 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :7:3: Syntax error: Operator expected
166 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :14:
167   Arithmetic: 'true' is not a function
168 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :20:
169   No permission to modify static procedure '(is)/2'
170 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :21:
171   No permission to modify static procedure '(is)/2'
172 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :24:3: Syntax error: Operator expected
173 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :26:
174   No permission to modify static procedure '(is)/2'
175 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Breeder_Reaction.pl
      :27:
```

```

176 No permission to modify static procedure '(is)/2'
177 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bridge Crossing.pl
      :6:
178 No permission to modify static procedure '(is)/2'
179 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bridge Crossing.pl
      :8:
180 Arithmetic: 'phenomenal' is not a function
181 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bridge Crossing.pl
      :15:
182 No permission to modify static procedure '(is)/2'
183 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Bright Islands.pl
      :8:3: Syntax error: Operator expected
184 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Brown John's Body.
      pl:11:3: Syntax error: Operator expected
185 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/By Earthlight.pl
      :4:1: Syntax error: End of file in quoted atom
186 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :1:3: Syntax error: Operator expected
187 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :3:3: Syntax error: Operator expected
188 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:5:
189 No permission to modify static procedure '(is)/2'
190 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:9:
191 No permission to modify static procedure '(is)/2'
192 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :10:3: Syntax error: Operator expected
193 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :16:3: Syntax error: Operator expected
194 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:19:
195 Arithmetic: 'possible' is not a function
196 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:20:
197 No permission to modify static procedure '(is)/2'
198 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :21:3: Syntax error: Operator expected
199 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:27:
200 No permission to modify static procedure '(is)/2'
201 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:29:
202 No permission to modify static procedure '(is)/2'

```

```
203 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :31:3: Syntax error: Operator expected
204 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl:34:
205   No permission to modify static procedure '(is)/2'
206 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cancer World.pl
      :37:3: Syntax error: Operator expected
207 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :2:3: Syntax error: Operator expected
208 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :5:3: Syntax error: Operator expected
209 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :8:
210   No permission to modify static procedure '(is)/2'
211 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :10:0: Syntax error: Operator expected
212 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :11:
213   No permission to modify static procedure '(is)/2'
214 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :14:
215   No permission to modify static procedure '(is)/2'
216 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :15:3: Syntax error: Operator expected
217 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :18:3: Syntax error: Operator expected
218 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :21:30: Syntax error: Operator expected
219 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :31:
220   No permission to modify static procedure '(is)/2'
221 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :37:
222   No permission to modify static procedure '(is)/2'
223 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :39:3: Syntax error: Operator expected
224 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :43:
225   No permission to modify static procedure '(is)/2'
```

```
226 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cause of Death.pl
      :44:3: Syntax error: Operator expected
227 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Chain Reaction.pl
      :2:13: Syntax error: Operator expected
228 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Chain Reaction.pl
      :6:
229   Arithmetic: 'atabank' is not a function
230 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Chain Reaction.pl
      :17:
231   No permission to modify static procedure '(is)/2'
232 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Circus.pl:8:
233   No permission to modify static procedure '(is)/2'
234 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Circus.pl:9:
235   Arithmetic: 'similar' is not a function
236 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Circus.pl:24:3:
      Syntax error: Operator expected
237 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Coming Attraction.
      pl:2:
238   No permission to modify static procedure '(is)/2'
239 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:3:
240   Arithmetic: 'fromVenus' is not a function
241 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:16:
242   No permission to modify static procedure '(is)/2'
243 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:20:
244   Arithmetic: 'rathercostly' is not a function
245 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:24:
246   No permission to modify static procedure '(is)/2'
247 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:31:
248   No permission to modify static procedure '(is)/2'
249 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Common Denominator.
      pl:33:
250   No permission to modify static procedure '(is)/2'
251 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Communication.pl
      :7:1: Syntax error: End of file in quoted atom
```

```
252 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Consignment.pl:5:
253   No permission to modify static procedure '(is)/2'
254 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Consignment.pl:8:
255   No permission to modify static procedure '(is)/2'
256 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Consignment.pl
      :11:3: Syntax error: Operator expected
257 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl:1:
258   No permission to modify static procedure '(is)/2'
259 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl:3:
260   No permission to modify static procedure '(is)/2'
261 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl
      :4:3: Syntax error: Operator expected
262 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl:7:
263   No permission to modify static procedure '(is)/2'
264 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl
      :9:3: Syntax error: Operator expected
265 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl
      :13:3: Syntax error: Operator expected
266 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl
      :15:
267   No permission to modify static procedure '(is)/2'
268 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Control_Group.pl
      :17:
269   Arithmetic: 'impossible' is not a function
270 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Crossroads_of
      Destiny.pl:4:9: Syntax error: Operator expected
271 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Crossroads_of
      Destiny.pl:9:
272   No permission to modify static procedure '(is)/2'
273 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Cry_from_a_Far
      Planet.pl:6:3: Syntax error: Operator expected
274 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dead_World.pl:8:
275   Arithmetic: 'absolute' is not a function
276 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dead_World.pl:11:3:
      Syntax error: Operator expected
277 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dead_World.pl:13:3:
      Syntax error: Operator expected
278 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dead_World.pl:19:3:
      Syntax error: Operator expected
```

```

279 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dearest Enemy.pl:1:
280   Arithmetic: 'self' is not a function
281 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dearest.pl:3:
282   Arithmetic: 'near' is not a function
283 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dearest.pl:4:
284   Arithmetic: 'near' is not a function
285 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dearest.pl:5:1:
      Syntax error: End of file in quoted atom
286 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Death Wish.pl:5:2:
      Syntax error: Unexpected end of file
287 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Death of a Spaceman
      .pl:4:3: Syntax error: Operator expected
288 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Diagnosis.pl:3:50:
      Syntax error: Operator expected
289 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Diagnosis.pl:13:33:
      Syntax error: Operator expected
290 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Diplomatic Immunity
      .pl:10:14: Syntax error: Operator expected
291 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Diplomatic Immunity
      .pl:12:14: Syntax error: Operator expected
292 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:2:
293   No permission to modify static procedure '(is)/2'
294 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:3:
295   No permission to modify static procedure '(is)/2'
296 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:7:
297   No permission to modify static procedure '(is)/2'
298 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:12:
299   No permission to modify static procedure '(is)/2'
300 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:16:3: Syntax error: Operator expected
301 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disaster Revisited.
      pl:19:1: Syntax error: End of file in quoted atom
302 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disqualified.pl
      :1:3: Syntax error: Operator expected
303 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disqualified.pl:4:

```

```

304 Arithmetic: 'amorethoroughinvestigation' is not a function
305 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disqualified.pl:8:
306 No permission to modify static procedure '(is)/2'
307 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disqualified.pl
      :13:9: Syntax error: Operator expected
308 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disqualified.pl
      :17:1: Syntax error: End of file in quoted atom
309 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disturbing_Sun.pl
      :3:
310 No permission to modify static procedure '(is)/2'
311 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Disturbing_Sun.pl
      :6:3: Syntax error: Operator expected
312 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dogfight--1973.pl
      :6:
313 No permission to modify static procedure '(is)/2'
314 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dogfight--1973.pl
      :10:
315 Arithmetic: 'closeenough' is not a function
316 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Dogfight--1973.pl
      :11:11: Syntax error: Operator expected
317 ERROR: /Users/anirudhprabhu/Dropbox/Work/Thesis/Rule_Files/Don't_Shoot.pl:2:1:
      Syntax error: End of file in quoted string
318 true.
319
320 ?-

```

Figure F.1: Detailed output loading all rule bases into the SWI-Prolog environment.