UNIVERSITÄT ZU LÜBECK

Can semantic similarities of words enhance common sense reasoning? A case study with prover E and SUMO.

Können semantische Ähnlichkeiten von Wörtern die Schlussfolgerungen des gesunden Menschenverstands verbessern? Eine Fallstudie mit Prover E und SUMO.

Bachelorarbeit

verfasst am
Institut für Software Engineering und Programming Languages

im Rahmen des Studiengangs
Informatik
der Universität zu Lübeck

vorgelegt von
Julian Britz

ausgegeben und betreut von
Prof. Dr. Diedrich Wolter

mit Unterstützung von
Moritz Bayerkuhnlein

Lübeck, den 06. Juli 2025

IM FOCUS DAS LEBEN

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Julian Britz

Zusammenfassung

Zusammenfassung

Abstract

Abstract

Contents

1 Introduction 1

1.1 Problem Statement 1

1.2 Hypothesis 2

1.3 Justification and Implications 3

1.4 Structure of this Work 3

2 Related Work 4

3 Preliminaries 6

3.1 Common Sense Reasoning 6

3.2 Word Embeddings 7

3.3 Grammars in Natural and Formal Languages 8

3.4 Automated Theorem Proving 9

3.5 Axiom Selection Strategies 9

3.6 Syntax-Based Selection: SInE 10

3.7 Vector-Based Selection 10

3.8 SeVEn: Sentence-Based Vector Encoding 11

3.9 Combining Vector-Based and Syntactic Selection 11

4 Experiments 13

4.1 Construction of White-Box Tests 13

4.2 Reengineering of Prior Work 14

4.3 Reengineering of Axiom Selection Results 14

4.4 Analysis of Reengineering 15

4.5 Expanding Axiom Selection Beyond Semantic and Syntactic Methods 18

4.6 Comparison of Large Language Models 20

4.7 Evaluation of Theorem Prover Configurations 21

5 Conclusion 23

6 Future work 24

Bibliography 25

1

Introduction

Automated theorem proving is a fundamental technique in formal logic and artificial intelligence. It enables the validation of logical statements based on a given set of axioms and inference rules. This thesis investigates whether incorporating frequently used axioms, referred to as core axioms, into a subset of the Adimen SUMO grammar improves proof success rates. The selected subset is determined by combining syntactic and semantic criteria to ensure both structural and conceptual relevance.

1.1 Problem Statement

Given a formal logical grammar, such as Adimen SUMO, and a conjecture C, the objective is to identify a subset of axioms that maximizes the probability of proving C. The process involves:

- Selecting axioms based on syntactic and semantic similarity, forming a set denoted as $\mathcal{A}_{\text{rel}}$.
- Testing whether adding frequently used axioms, $\mathcal{A}_{\text{core}}$, improves proof success.

Even though there is no direct mapping between natural and logical language, we assume, that there is an entanglement between them:

$$\mathcal{L}_{\text{NL}} \bowtie \mathcal{L}_{\text{logic}} \tag{1.1}$$

Arguments for this assertion are that logical langugage is human made and therefore inevitably connected to human and consequently natural language. Also grammers like Adimen SUMO are partially derived by lexical-semantic networks that capture relationships between words of natural language. Making use of this entanglement the selection process of theorem proovers can be optimized by reducing the number of relevant axioms. Theorem prover E follows a refutational proof procedure using the superposition calculus (Schulz, Cruanes, and Vukmirovic, 2019). The proof search is based on a modified given-clause algorithm, consisting of several key steps:

1. Initialization: The input problem is converted into clausal normal form (CNF) and divided into two sets: unprocessed clauses U and processed clauses P.

2. Clause Selection: A clause g is selected from U based on heuristics, referred to as the given clause, and is then simplified against P.
3. Simplification: The given clause undergoes forward simplification using processed clauses. This step removes redundant literals, applies rewriting rules, and ensures that terms are arranged in a canonical order.
4. Inference: The prover applies the superposition calculus, combining g with clauses in P to generate new clauses. These new clauses are simplified and added to U if they are neither trivial nor redundant.
5. Saturation or Termination: The process stops when:
   - The empty clause $\perp$ is derived, confirming the proof.
   - The clause set is saturated, meaning no further inferences can be made.
   - A predefined resource limit, such as time or memory, is reached.

Axioms play a crucial role in simplification. While the prover cannot arbitrarily discard axioms, as they form the foundation of logical reasoning, they are essential for clause rewriting and subsumption.

1.2 Hypothesis

To test whether adding frequently used axioms improves proof success, we define the key elements of the problem:

- $\mathcal{A}$ represents the full set of axioms available in Adimen SUMO.
- C is the conjecture we want to prove.
- d is a function that measures how similar an axiom is to the conjecture, based on both syntax and meaning.
- $\mathcal{A}_{\text{rel}}$ is the set of k axioms that are most similar to the conjecture according to d:

$$\mathcal{A}_{\text{rel}} = \{A \in \mathcal{A} \mid A \text{ is among the } k \text{ closest axioms to } C \text{ w.r.t. } d\} \tag{1.2}$$

- $\mathcal{A}_{\text{core}}$ is a collection of axioms that have been used frequently in past proofs.
- The enhanced axiom set, $\mathcal{A}_{\text{enh}}$, is created by combining the most similar axioms with the frequently used ones:

$$\mathcal{A}_{\text{enh}} = \mathcal{A}_{\text{rel}} \cup \mathcal{A}_{\text{core}} \tag{1.3}$$

- $P(T, \mathcal{A}', C)$ is a function that checks if the theorem prover T is able to prove C using a given set of axioms $\mathcal{A}'$. If the proof is successful, the function returns true; otherwise, it returns false.

The hypothesis we test is:

$$\forall C \in \mathcal{C}, \quad \Pr(P(T, \mathcal{A}_{\text{enh}}, C) = \text{True}) > \Pr(P(T, \mathcal{A}_{\text{rel}}, C) = \text{True}) \tag{1.4}$$

In other words, the likelihood of proving a conjecture is higher when frequently used axioms are included along with the most similar ones.

1.3 Justification and Implications

Adimen SUMO's grammar is closely connected to natural language structures. This means that selecting axioms based on semantic embeddings, such as SBERT, helps capture underlying logical dependencies that might not be immediately apparent. Frequently used axioms act as structural anchors within this system, making it easier for the theorem prover to navigate the proof search space and discover relevant inference steps. By combining the most relevant axioms with those that have been frequently used in past proofs, the theorem prover benefits from both semantic alignment and structural consistency. This increases the overall proof success rate, as the prover has access to axioms that not only appear relevant but have also proven useful in previous reasoning tasks.

This thesis tests this hypothesis through empirical experiments, measuring proof success rates across different axiom selection strategies. The goal is to determine how much incorporating core axioms improves theorem proving performance.

1.4 Structure of this Work

This thesis is structured into several chapters, each addressing different aspects of the research problem and methodology. It begins with an overview of existing research, followed by the theoretical foundations, the proposed approach, experimental evaluations, and concluding discussions.

Chapter 2 provides a review of related work. It covers previous research on theorem proving and axiom selection techniques, as well as how semantic similarity has been applied in logical reasoning.

Chapter 3 introduces fundamental concepts necessary to understand the core of this research. It discusses topics such as common sense reasoning, word embeddings, and theorem proving. Additionally, it explains the role of formal grammars and axiom selection in automated reasoning.

Chapter 4 focuses on the experimental setup and evaluation. It outlines the benchmark datasets, describes the evaluation metrics, and presents the results of different axiom selection strategies. The impact of incorporating core axioms into theorem proving is analyzed based on empirical findings.

Chapter 5 summarizes the main contributions of this work and highlights key insights gained from the experiments. It also discusses potential future directions and remaining challenges in axiom selection and automated reasoning.

Finally, Chapter 6 explores ideas for future research, suggesting possible extensions and refinements to the proposed approach.

Each chapter builds upon the previous ones, providing a structured progression from theoretical background to practical implementation and evaluation.

2

Related Work

A significant part of this thesis builds upon the work of Schon et al. (Jakobs and Schon, 2024), which explores context-specific axiom selection using large language models. Traditional axiom selection methods primarily rely on syntactic properties, often ignoring the meaning embedded in symbol names. Schon proposes an alternative approach by leveraging large language models to determine relevant axioms based on contextual similarity. This method aligns axiom selection with the goal's context, improving performance in commonsense reasoning tasks. Experiments show that this approach outperforms purely syntactic selection methods, demonstrating the potential of semantic similarity for guiding theorem provers.

The findings of Schon's work play a crucial role in this thesis, as a major aspect of the proposed approach involves using semantic embeddings for axiom selection. However, while Schon focuses on direct context-based selection, this thesis investigates the impact of incorporating frequently used axioms into the selection process. The goal is to determine whether combining core axioms with those chosen based on semantic similarity enhances proof success rates.

Beyond Schon's work, several other studies have contributed to axiom selection and theorem proving. Adimen-SUMO, developed by Alvez et al. (Alvez, Lucio, and Rigau, 2014), is a reengineered version of the SUMO ontology designed to improve compatibility with first-order theorem provers. The restructuring ensures that axioms are formulated in a way that facilitates inference, reducing ambiguities present in the original SUMO. This ontology serves as a key resource in evaluating theorem provers, including those used in this thesis.

Syntactic axiom selection has also been widely explored. Hoder and Voronkov (Hoder and Voronkov, 2011) introduced SInE, a system that selects axioms based on their syntactic relevance to the conjecture. By iteratively choosing axioms that introduce symbols appearing in the conjecture, SInE reduces the search space, making large knowledge bases more manageable. Similarly, Roederer et al.

Another relevant contribution comes from Alvez et al. (Alvez et al., 2017), who introduced a framework for systematic white-box testing of first-order logic ontologies. Their work focuses on ensuring logical consistency and eliminating redundant or contradictory axioms. Automated testing techniques from this study provide insights into ontology refinement, which is relevant when evaluating the quality of selected axioms in theorem proving.

In summary, Schon's research serves as the primary foundation for this thesis, providing a framework for semantic axiom selection that is further extended through the integration of frequently used axioms. Other related studies on syntactic axiom selection, ontology structuring, and logical consistency contribute valuable insights, supporting the refinement of the proposed approach.

3

Preliminaries

This chapter introduces fundamental concepts relevant to this work. It covers common sense reasoning, word embeddings, formal grammars, automated theorem proving, and axiom selection strategies. These topics form the theoretical foundation for the experiments and analyses presented in later chapters.

3.1 Common Sense Reasoning

Common sense reasoning is an essential part of human cognition, allowing individuals to understand everyday situations, infer unstated knowledge, and make reasonable decisions even when information is incomplete. Automated reasoning systems attempt to replicate this ability to improve inference and decision-making.

Common sense reasoning involves recognizing unstated premises in logical arguments, understanding cause-effect relationships, and applying heuristics to solve problems under uncertainty. It plays a key role in artificial intelligence, particularly in natural language understanding, where context is crucial for interpreting meaning, and in automated reasoning, where structured logical frameworks help infer conclusions from available premises. Applications in law, medicine, and finance further highlight the importance of integrating knowledge-based reasoning for interpretability and reliability. Despite advancements in AI, deep learning models struggle with incorporating structured and generalizable knowledge, making common sense reasoning an ongoing challenge.

Interpreting a sentence often involves more than just processing its words—it requires incorporating background knowledge to make sense of implicit information. Take, for example, the Winograd Schema involving John and Billy viewing a stage (Levesque, Davis, and Morgenstern, 2012). While the sentence itself is simple, understanding it correctly depends on a series of assumptions that humans naturally make when constructing a mental model of the situation (Bayerkuhnlein, 2023).

A stage is something meant to be viewed. In a typical setting, viewers remain stationary, and obstructions can block their line of sight. Humans are physical entities, meaning they cannot see through solid objects. Both John and Billy are human names, so the sentence suggests a scene where one person is obstructing the other's view. The interpretation of who is blocking whom depends on the additional context provided by the sentence, but the reasoning process remains the same: we piece together knowledge about how the world works to arrive at a coherent understanding.

These intuitive assumptions form the foundation of common sense reasoning. The process aligns with the theory of mental models, which suggests that humans construct internal representations of a scenario to evaluate possible interpretations (Johnson-Laird, 1989). In the case of ambiguous pronouns, this internal model helps determine which assignment makes the most logical sense. This ability to build structured mental representations is what allows humans to effortlessly resolve ambiguities that challenge AI systems.

3.2 Word Embeddings

Word embeddings represent words as numerical vectors in a continuous space, capturing semantic relationships based on their contextual usage. Unlike traditional representations such as one-hot encoding, embeddings enable words with similar meanings to be positioned closer together in the vector space, allowing models to generalize across linguistic variations.

Different techniques have been developed to generate word embeddings, each with unique properties and advantages. Word2Vec (Mikolov et al., 2013) is one of the foundational models, introducing two architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a word based on its surrounding context, while Skip-gram does the inverse, predicting context words given a target word. This approach efficiently captures both syntactic and semantic relationships. FastText (Bojanowski et al., 2017) extends Word2Vec by incorporating subword information, representing words as bags of character n-grams. This allows it to handle rare words and out-of-vocabulary terms better than standard word-based models. GloVe (Pennington, Socher, and Manning, 2014), in contrast, constructs embeddings by factorizing a word co-occurrence matrix, capturing both local and global semantic information.

The effectiveness of word embeddings depends on several factors, including the size of the context window, the type of training data, and the inclusion of subword information.

Beyond standard word embeddings, contextual embeddings such as BERT and GPT (Devlin et al., 2019) further refine representation by considering the full sentence structure, allowing word meanings to dynamically change based on context. This advancement is particularly relevant in cases where words have multiple meanings depending on their use in different sentences.

In this work, word embeddings are leveraged to determine the semantic similarity between axioms and conjectures. By using pre-trained models like Sentence-BERT (SBERT), it becomes possible to enhance axiom selection in theorem proving, ensuring that the selected axioms are not only structurally relevant but also semantically meaningful.

3.3 Grammars in Natural and Formal Languages

Grammars define the structural rules of a language, whether it is a natural language like English or a formal system used in logic. In theorem proving, grammars are essential for structuring logical statements in a way that machines can interpret and manipulate.

Different types of grammars serve various purposes. Context-free grammars (CFGs) define syntactic structures in natural language processing and programming languages. First-order logic (FOL) grammars represent logical expressions using predicates, functions, and quantifiers. Ontology-based grammars, such as SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001), combine linguistic and logical structures to improve knowledge representation.

Grammars are crucial in theorem proving because they enforce logical consistency, allow automated systems to process logical expressions systematically, and provide a structured representation of knowledge in formal systems like SUMO and Adimen-SUMO. Adimen-SUMO is written in Thousands of Problems for Theorem Provers (TPTP) Syntax, a standardized language designed for expressing logical problems in automated reasoning (Alvez, Lucio, and Rigau, 2014). TPTP, originally developed to facilitate the benchmarking and comparison of theorem provers, provides a well-defined structure for first-order logic expressions, ensuring compatibility across different proving systems. By using TPTP, Adimen-SUMO maintains a formal and machine-readable representation of axioms, allowing theorem provers to efficiently process and reason over its logical content.

In TPTP syntax, axioms are typically expressed in First-Order Form (FOF), which represents first-order logic statements in a structured and readable format. The FOF format explicitly defines logical components such as axioms, conjectures, and definitions, ensuring clarity in automated reasoning tasks. An example of an axiom in Adimen-SUMO written in FOF syntax is:

Listing 3.1: Axiom example Adimen-SUMO

```
fof(predefinitionsA7, axiom,
    (![X]:
        (
            p__d__subclass(X, X)
        )
    )
).
```

This axiom states that for all entities X, X is always a subclass of itself. In natural language, this expresses the reflexive property of the subclass relation, meaning that every class is considered a subclass of itself. This kind of axiom is fundamental in ontology-based reasoning, as it helps maintain consistency within hierarchical structures.

3.4 Automated Theorem Proving

Automated theorem proving aims to determine whether a given conjecture follows logically from a set of axioms. It is a key component of formal logic and artificial intelligence.

Theorem provers rely on different proof strategies. Logical inference derives new statements from existing ones using formal rules. Resolution is a proof technique that derives contradictions, commonly used in first-order logic provers. Refutation works by attempting to show that a conjecture holds by deriving a contradiction from its negation.

Prover E is a widely used theorem prover based on the superposition calculus (Schulz, Cruanes, and Vukmirovic, 2019). Another prominent theorem prover is Vampire, a first-order logic prover known for its efficient implementation of resolution and superposition techniques (**Riazanov2002**). Vampire applies a range of strategies, including term indexing and subsumption algorithms, to optimize proof search.

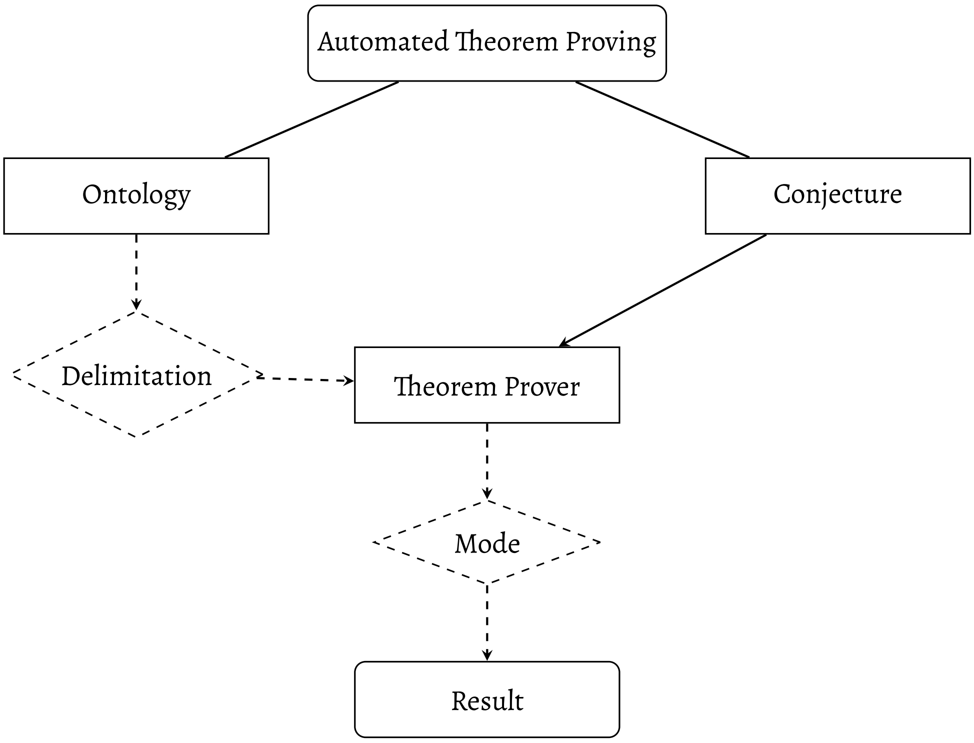

Figure 3.2: Automated Theorem Proving

3.5 Axiom Selection Strategies

Theorem provers often struggle with large search spaces, making axiom selection a critical challenge. The process of selecting axioms influences proof efficiency and success rates.

3.6 Syntax-Based Selection: SInE

The SInE selection strategy (Hoder and Voronkov, 2011) is a trigger-based approach designed to efficiently select axioms in automated theorem proving. Instead of selecting axioms based on fixed rules, it dynamically determines relevance using a trigger relation. This relation ensures that an axiom is only selected if at least one of its symbols appears less frequently than or as frequently as all other symbols in the axiom. This prevents highly common symbols from triggering every axiom they appear in, reducing unnecessary selections.

The selection process starts by identifying symbols that appear in the conjecture to be proven. These symbols are considered directly triggered. If a triggered symbol appears in an axiom and satisfies the trigger relation, the axiom is selected, and its other symbols become triggered in the next selection step. This process continues recursively up to a defined depth k, ensuring that only the most relevant axioms are included.

One limitation of this approach arises when an axiom contains symbols with nearly equal occurrence counts. In such cases, only the least frequent symbol can trigger the axiom, potentially excluding useful axioms. To address this, a benevolence parameter b is introduced, allowing symbols that occur up to b times more frequently than the least common symbol to also act as triggers. This adjustment provides more flexibility in axiom selection.

3.7 Vector-Based Selection

Syntax-based selection methods like SInE do not consider the meaning of symbol names, which can lead to the loss of useful information. This is particularly problematic in commonsense reasoning, where symbol names often carry semantic meaning that could improve selection. To address this, vector-based selection methods leverage word embeddings to quantify the semantic similarity between axioms and the goal.

As seen above, word embeddings are a common technique in natural language processing, based on the distributional hypothesis (Miller and Charles, 1991). This hypothesis states that words appearing in similar contexts tend to have similar meanings. By training neural networks on large text corpora, words are mapped into a continuous vector space, where their geometric relationships capture semantic similarity. Cosine similarity is commonly used to measure the closeness of two word vectors, with values ranging from -1 (completely dissimilar) to 1 (identical).

Definition 3.1. Let $u, v \in \mathbb{R}^n$ be non-zero vectors. The cosine similarity of u and v is defined as:

$$\cos\mathrm{sim}(u, v) = \frac{u \cdot v}{||u||||v||} \tag{3.1}$$

Schon (Schon, 2023) introduces a statistical approach for axiom selection that relies on word embeddings to represent axioms as vectors. The idea is to encode axioms into a continuous vector space where semantically similar axioms are mapped closer together. This is done in a preprocessing step, where each axiom in the knowledge base is transformed into a vector representation using a word embedding model.

When selecting axioms for a given goal G, the goal is vectorized in the same way as the axioms in the knowledge base. The selection process then identifies the n axioms whose vector representations are most similar to that of G. The similarity between vectors is measured using cosine similarity, which quantifies how closely two vectors align regardless of their magnitude.

Formally, given a knowledge base KB, a goal G, and a vectorization function f that maps axioms to $\mathbb{R}^n$, the vector-based selection process selects the n axioms in KB that have the highest cosine similarity to G. The selected axioms form a subset of KB where no other axiom outside this subset has a higher similarity to G than the least similar axiom within the selected set.

This approach ensures that axioms most relevant to the goal, based on their semantic similarity, are prioritized during the selection process.

3.8 SeVEn: Sentence-Based Vector Encoding

Sentence embeddings function similarly to word embeddings, as they are also generated by training neural networks on large text corpora. However, while word embeddings focus on individual words, sentence embeddings capture the meaning of entire sentences or even full documents (Kiros et al., 2015). This ability makes them particularly useful for encoding more complex relationships between axioms.

Following this idea, the SeVEn approach builds upon vector-based selection by modifying how axioms are represented. Instead of encoding individual symbol names, each axiom is first translated into a sentence that captures its semantic meaning. This sentence is then encoded into a vector using a sentence embedding model. The overall process remains similar to standard vector-based selection, but the vectorization method is adapted to work at the sentence level. Given a knowledge base KB, where each axiom A is transformed into a sentence S using a function t, the sentence embedding function f then maps S into a vector representation $v_S(A)$. The full sentence-based vector representation of the knowledge base is denoted as $V_S(KB)$, where each axiom is now represented as a sentence embedding.

This adaptation allows SeVEn to leverage the richer semantic information captured in sentence embeddings while maintaining the core structure of vector-based selection.

3.9 Combining Vector-Based and Syntactic Selection

.

To address this issue, hybrid selection strategies combine statistical and syntactic approaches. An example of this is Similarity SInE (Furbach, Kramer, and Schon, 2019), which extends the original SInE selection by allowing not just the least frequent symbol to trigger an axiom but also symbols with a high enough cosine similarity to the trigger.

This modification ensures that semantically related axioms can still be included in the selection process.

Building on this idea, (Schon, 2023) introduced a hybrid approach that integrates SeVEn with the syntactic selection method of SInE. The process starts by expanding the goal with axioms selected through SeVEn. Then, SInE is applied to this extended set of conjectures, selecting additional axioms based on its trigger-based mechanism.

Using formal notation, let $\mathrm{sine}(KB, \{A_1, \ldots, A_n\})$ represent the axioms selected by SInE for a set of conjectures, and let $\mathrm{seven}(KB, G)$ denote the axioms selected by SeVEn for a goal G. The combined selection strategy is then defined as:

Definition 3.2.

$$\mathrm{union\_select}(KB, G) = \mathrm{sine}(KB, G \cup \mathrm{seven}(KB, G)) \tag{3.2}$$

This formulation allows for flexibility in axiom selection, as different selection strategies can be incorporated depending on the specific needs of the reasoning task.

4

Experiments

This chapter presents the experimental setup and evaluation of different axiom selection strategies in the context of automated theorem proving. The experiments focus on white-box testing, reengineering of prior work, comparative analysis of selection methods, and performance assessments across different theorem provers.

4.1 Construction of White-Box Tests

The methodology for constructing white-box tests follows the approach described by Alvez (Alvez et al., 2017). The tests are designed to assess the logical consistency and inference capabilities of theorem provers by introducing controlled variations in axiom selection.

In the context of Adimen SUMO, falsity tests are constructed by taking specific statements or conditions and applying negation. These negated statements form the basis for falsity tests, which are essential for evaluating logical proof systems. A theorem prover such as Prover E processes these tests through a method called refutation. The goal of refutation is to show that the original statements are inconsistent or false. This is done by attempting to derive a contradiction from the negated statements. If a contradiction is found, it confirms that the original, non-negated statements cannot all be true at the same time.

An example of a white-box falsity test in Adimen-SUMO is the following:

Listing 4.1: WhiteboxTruthTest example

```
fof( whiteBoxTruthTest2824, conjecture,
    (?[MORPH]:
        (
            ~ (
                p__d__instance(MORPH,c__Morpheme)
            )
        )
    )
).
```

The expression states that there exists at least one entity, represented by the variable $MORPH$, for which the predicate $p\_\_d\_\_instance(MORPH, c\_\_Morpheme)$ does not hold. In simpler terms, it asserts that at least one entity is not an instance of the class "Morpheme."

By negating the statement that all instances belong to the class "Morpheme," this test challenges the theorem prover to derive a contradiction if the ontology implies that every entity must belong to this class. If a contradiction is found, it confirms that the original classification of "Morpheme" is inconsistent, helping to assess the reasoning capabilities of the prover.

4.2 Reengineering of Prior Work

4.3 Reengineering of Axiom Selection Results

To evaluate different axiom selection techniques, this work reengineers the results presented in (Jakobs and Schon, 2024). The primary goal is to validate previous findings while assessing the impact of alternative selection methods. The evaluation compares three selection approaches: syntactic selection using SInE, semantic selection based on vector embeddings, and a hybrid union approach combining both methods.

Each method is tested under the same conditions using a randomly selected subset of 1,000 white-box truth tests from Adimen-SUMO. The theorem prover E is used with a fixed configuration, ensuring consistent input formatting and a time limit of 15 seconds per proof attempt:

Listing 4.2: Prover E configuration

```
eprover ––tstp–format ––soft–cpu–limit=15
```

For syntactic selection, the SInE strategy is applied with a benevolence parameter of 3 and a recursion depth of 2. In contrast to prior work, the number of selected axioms in the semantic analysis is increased to 1,500, which is three times higher than the previous maximum. This adjustment is made to align the number of selected axioms with the average output of the SInE strategy, ensuring comparability between the syntactic and semantic runs.

The hybrid union approach, which combines vector-based and syntactic selection, is configured as follows:

- The sentence embedding model used is the pre-trained all-MiniLM-L6-v2 from SentenceTransformer, featuring 384 dimensions and a model size of 80MB. It has been trained on a large dataset of over 1 billion training pairs and is optimized for diverse NLP tasks (**Reimers2019**; **Sentencetransformers2019**).
- The number of closest axioms n selected for each goal G is set to 160.
- The SInE parameters remain the same as in the syntactic selection: benevolence = 3, recursion depth = 2.

Since the union approach applies the SInE strategy to a broader set of initial axioms than the purely syntactic selection, the final number of selected axioms is also higher, averaging around 1,800. This increase reflects the integration of both semantic and syntactic relevance in axiom selection, aiming to enhance the theorem prover's ability to find proofs efficiently.

The prover results are categorized into:

- *Timeout:* Instances where the automated prover exceeded the allocated computational resources without deriving a proof.
- *Proof Found:* Instances where the prover successfully established a proof within the given constraints.
- *Gave Up:* Instances where the prover terminated prematurely without reaching a conclusive result.

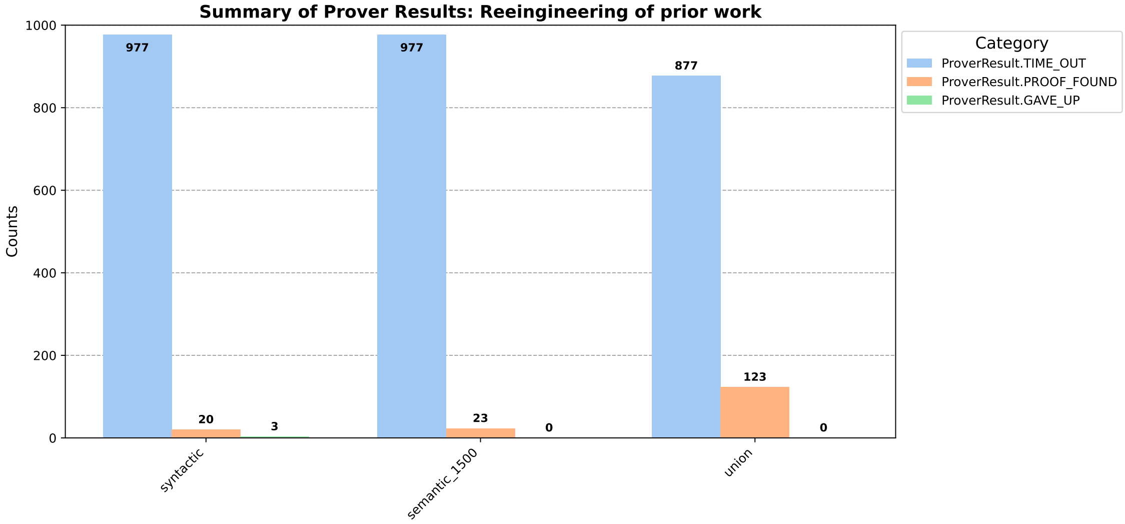

Figure 4.3: Reengineering of results from (Jakobs and Schon, 2024)

The results in 4.3 correspond to the prior investigations. While a purely syntactic and purely semantic approach can only find few proofs in the given time, the combination can achieve better results. A significant portion of test cases resulted in timeouts or failed proof attempts. But also for the combination the overall success rate remained low. The findings suggest that relying solely on semantic similarity in combination with a syntactical selection precedure is not sufficient for effective axiom selection. Future experiments should investigate whether incorporating frequently used core axioms can enhance proof success rates. Additionally, exploring the relationship between axiom complexity and proof outcomes may offer deeper insights into theorem proving behavior, ultimately leading to more refined selection strategies.

4.4 Analysis of Reengineering

To get an indication and possibly recognize patterns for which conjectures the prover could find proofs, the results of the previous union run are examined. To understand the factors affecting the success of proof search in theorem proving, several characteristics of the conjectures are examined. Specifically, the influence of the number of variables, the presence of logical operators, and the overall character count in a conjecture are analyzed. These elements contribute to the complexity of the search space and potentially impact the likelihood of finding a proof.

Since each variable expands the number of possible substitutions and inferences the prover must consider, a higher number of variables generally results in a larger search space. It is expected that conjectures with fewer variables are easier to prove, as the prover has to evaluate fewer possible cases.

First-order logic relies on specific symbols as logical operators, each contributing to the structural complexity of a conjecture. The most relevant symbols include:

- | (disjunction) – represents logical OR.
- $ (conjunction) – represents logical AND.
- ? (variable indicator) – denotes the existence of a variable.
- ! (universal quantifier) – states that a property holds for all instances of a variable.
- ~ (negation) – expresses logical NOT.

A higher occurrence of these symbols increases the structural complexity of a conjecture. More complex logical expressions require additional inference steps, making it more challenging for the theorem prover to find a proof. Therefore, it is expected that conjectures with fewer logical operators are more likely to be proven within a limited time.

Lastly, the overall character count of a conjecture is analyzed as a general measure of complexity. Longer conjectures tend to introduce more structural elements, increasing the reasoning effort required by the prover. By examining how the length of a conjecture correlates with proof success rates, it is possible to determine whether shorter conjectures are indeed easier to prove.

The following presents the results of these analyses, comparing proof success rates based on these complexity factors. The implications of these findings are then discussed in detail, highlighting how different structural properties influence theorem proving performance.

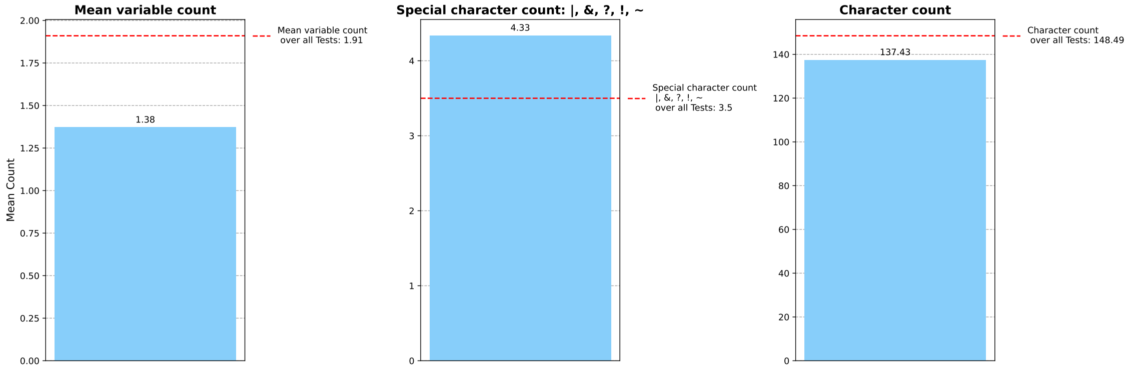

Figure 4.4: Analysis of Reengineering

The results confirm that all examined factors influence the success rate of theorem proving as expected. Conjectures with fewer variables consistently yielded higher proof success rates, supporting the assumption that an increase in the number of variables expands the search space, making proof discovery more challenging.

Similarly, the impact of logical operators aligns with expectations. Conjectures containing a higher number of disjunctions, conjunctions, quantifiers, and negations were significantly harder to prove.

The general character count of a conjecture also showed a correlation with proof success. Conjectures with a lower character count were proven more frequently, reinforcing the idea that shorter conjectures introduce fewer structural elements that could complicate reasoning. While character count alone is not a direct measure of complexity, it serves as a useful indicator of the structural and syntactic complexity of a logical statement.

Next to analyze how the cosin similarity of the 160 chosen axioms differs between those proofs, for which a proof could be found and those not.

Next, the cosine similarity of the 160 selected axioms is analyzed to compare cases where a proof was successfully found to those where no proof could be derived. This evaluation aims to determine whether higher semantic similarity between the selected axioms and the conjecture correlates with a higher proof success rate.

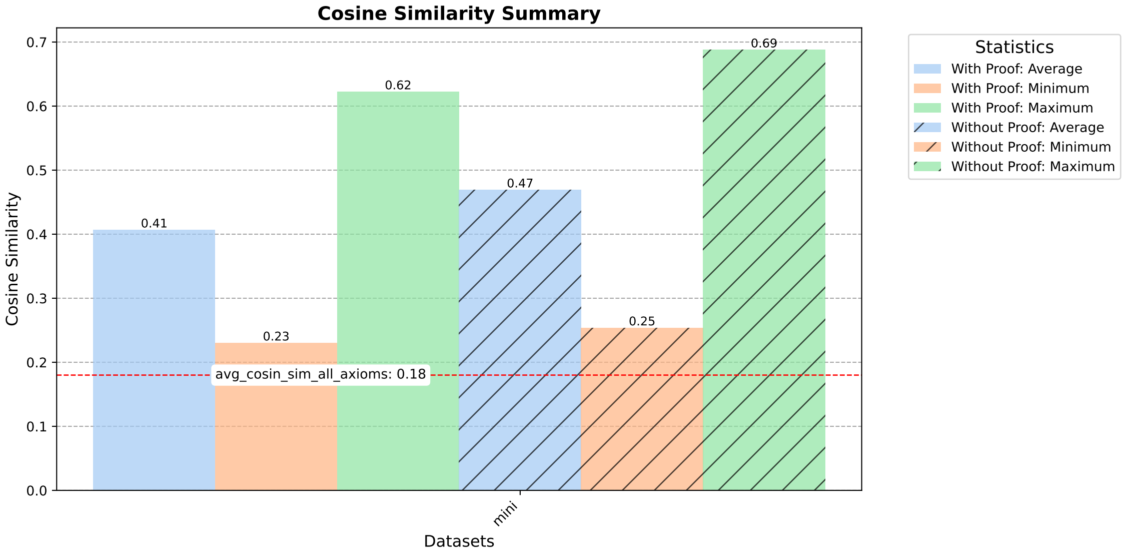

Figure 4.5: Cosine similarity distribution

Unexpectedly, the average cosine similarity between the conjecture and the 160 nearest axioms is slightly lower in cases where a proof was found compared to those where no proof was derived. This suggests that an excessively high cosine similarity may restrict the search space too much, making the selection of axioms overly specific and limiting the prover's ability to construct a valid proof. This indicates that proving a conjecture may require bridging different concepts rather than simply refining closely related axioms.

4.5 Expanding Axiom Selection Beyond Semantic and Syntactic Methods

As previously observed, the pure combination of syntactic and semantic selection methods does not cover a sufficiently broad range of proofs. The results show that conjectures without a successful proof tend to have a higher cosine similarity to their 160 nearest axioms. This suggests that relying solely on semantic and syntactic similarity is not enough. A too-specific selection of axioms can lead to missing essential axioms, ultimately preventing the prover from finding a proof.

Since theorem provers, as described in 1, rely on simplification steps using available axioms, it becomes evident that additional axioms enabling these simplifications are required. Without such axioms, the prover lacks the necessary transformation rules to progress toward a proof.

To identify the core axioms, Prover E was executed in auto mode across all 8,010 white-box truth tests. In this mode, the prover automatically determines the optimal SInE parameters and applies a search heuristic best suited for each conjecture. After running all tests, the 25 axioms that appeared most frequently in successful proofs were selected as core axioms.

Listing 4.6: Prover E configuration

```
eprover --auto --tstp-format --soft-cpu-limit=15
```

The choice of these axioms is based on their practical relevance in proof construction. Since they are consistently used in successful proofs, they likely provide essential logical connections and simplifications that improve the theorem prover's efficiency.

An example of a core axiom from Adimen-SUMO that facilitates inference is:

Listing 4.7: Example core axiom

```
fof( mergeA594, axiom,
    (![REL]:
        (
            (
                p__d__instance(REL,c__BinaryRelation)
            )
            =>
            (
                (
                    p__d__instance(REL,c__IrreflexiveRelation)
                )
                <=>
                (
                    (![INST]:
                        (
                            ~ (
                                p__d__holds3(REL,INST,INST)
                            )
                        )
                    )
                )
            )
        )
    )
).
```

This axiom defines the irreflexive property of certain binary relations. It states that if a relation REL is classified as a binary relation, then it is an irreflexive relation if and only if there exists no instance INST for which REL holds between INST and itself. In simpler terms, this means that for a relation to be irreflexive, it must never relate an entity to itself.

Such an axiom is fundamental in reasoning about relational structures, as many logical proofs rely on distinguishing between reflexive and irreflexive relations. Without it, the prover may struggle to validate properties of relations that are essential for defining hierarchical and ordering constraints. Including these core axioms in the selection process ensures that theorem proving is not only driven by semantic and syntactic similarity but also by foundational logical principles necessary for accurate inference.

To increase the likelihood of selecting relevant core axioms without significantly expanding their number, a SInE strategy was applied to the 25 core axioms with a benevolence value of 1 and a recursion depth of 1. The axioms selected through this process were then combined with those obtained from the union-based selection method, ensuring that both frequently used and structurally relevant axioms were included. Applying this on the same 1.000 WhiteboxTruthTests as before, 286 proofs could be found.

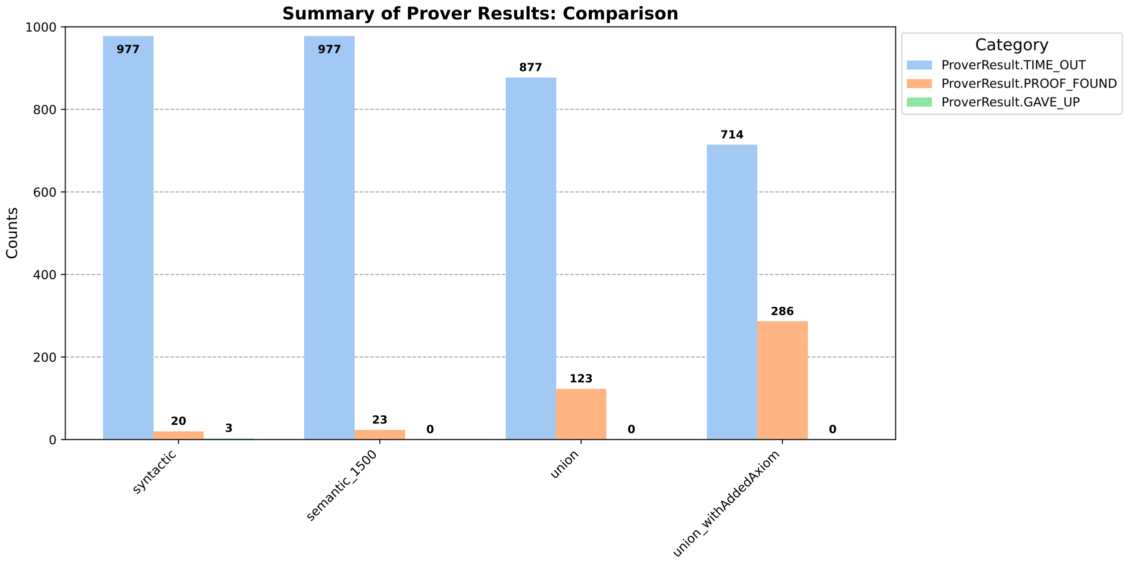

Figure 4.8: Summary of prover results with core axioms

The union run successfully found 123 proofs. After adding the core axioms, the number of proofs increased by 163, resulting in a total of 286 proofs. A closer analysis showed that all 123 proofs from the union run were a subset of the proofs found when core axioms were included.

$$P_{\text{union}} \subseteq P_{\text{core}} \cup P_{\text{union}} \tag{4.1}$$

where P_{union} represents the set of proofs found in the union run and P_{core} represents the set of proofs found after adding core axioms.

This suggests that adding core axioms plays a crucial role in enabling the prover to find proofs that would otherwise remain undiscovered.

4.6 Comparison of Large Language Models

The performance of different large language models in axiom selection is assessed.

- The results of theorem proving with different models are analyzed.

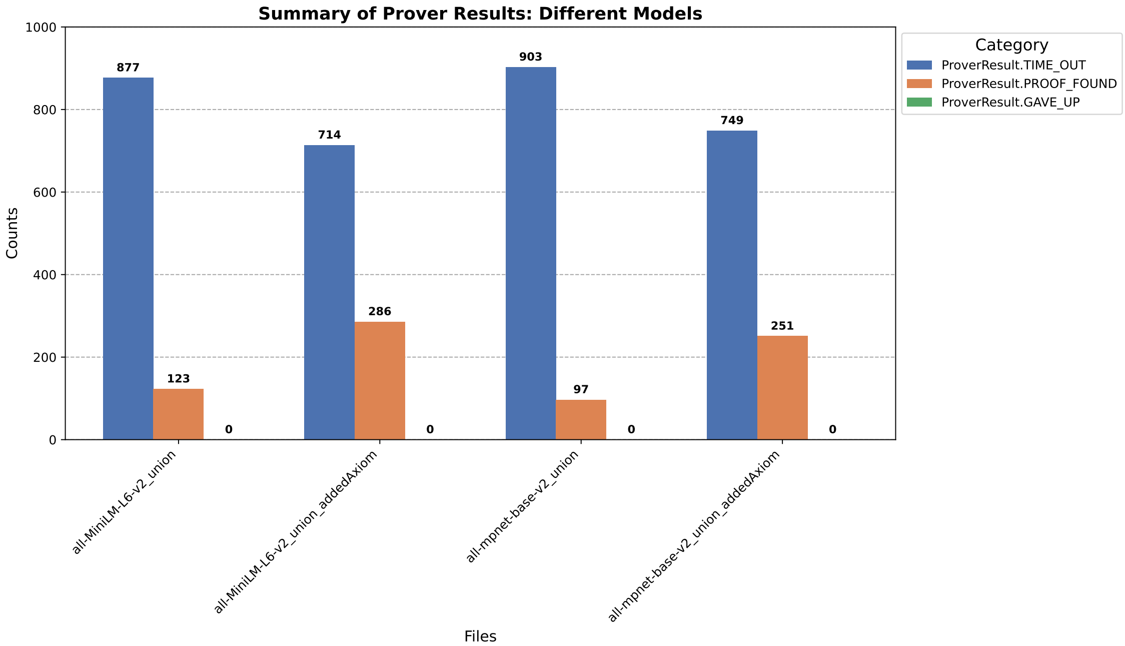

Figure 4.9: Summary of prover results with different models

- The cosine similarity of selected axioms is compared.

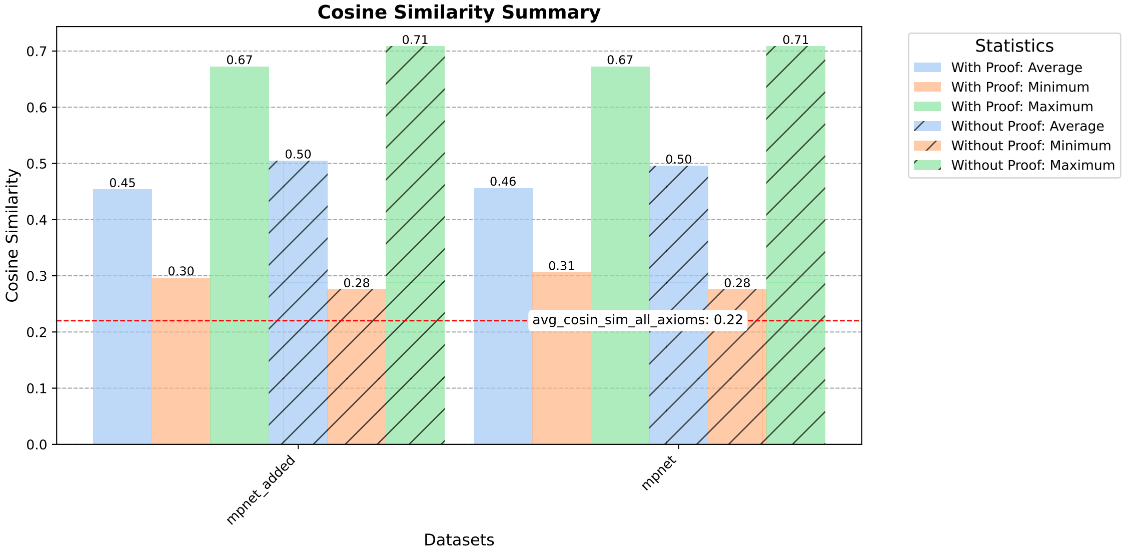

Figure 4.10: Cosine similarity in different models

4.7 Evaluation of Theorem Prover Configurations

The effect of different theorem prover configurations is analyzed.

- The performance of Satauto mode is examined.

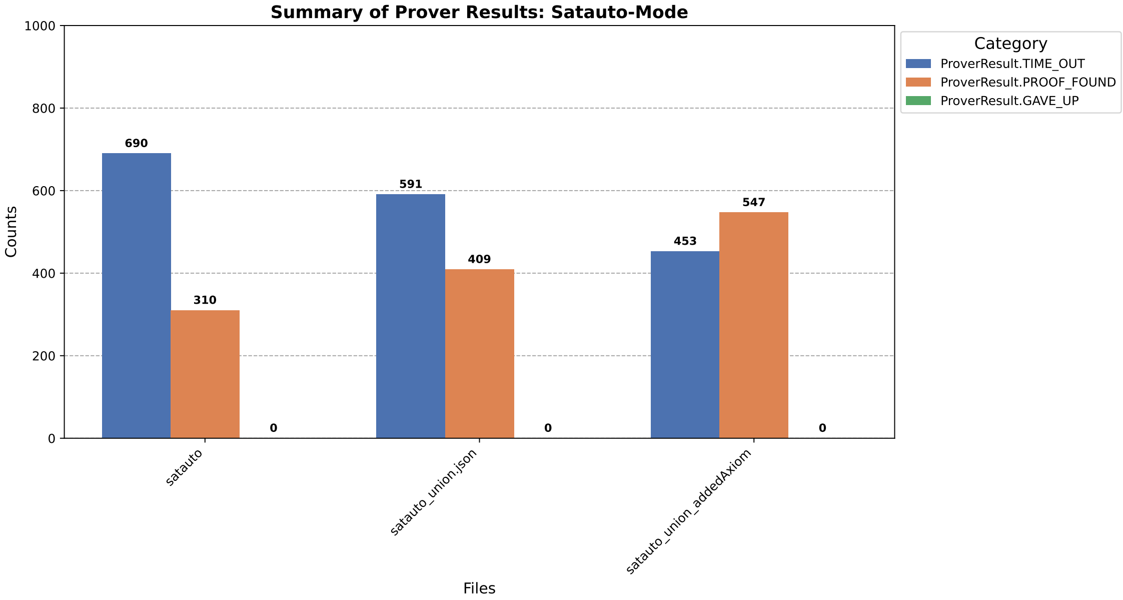

Figure 4.11: Summary of prover results in Satauto mode

- The effect of auto mode, which uses SInE and search heuristics, is assessed.

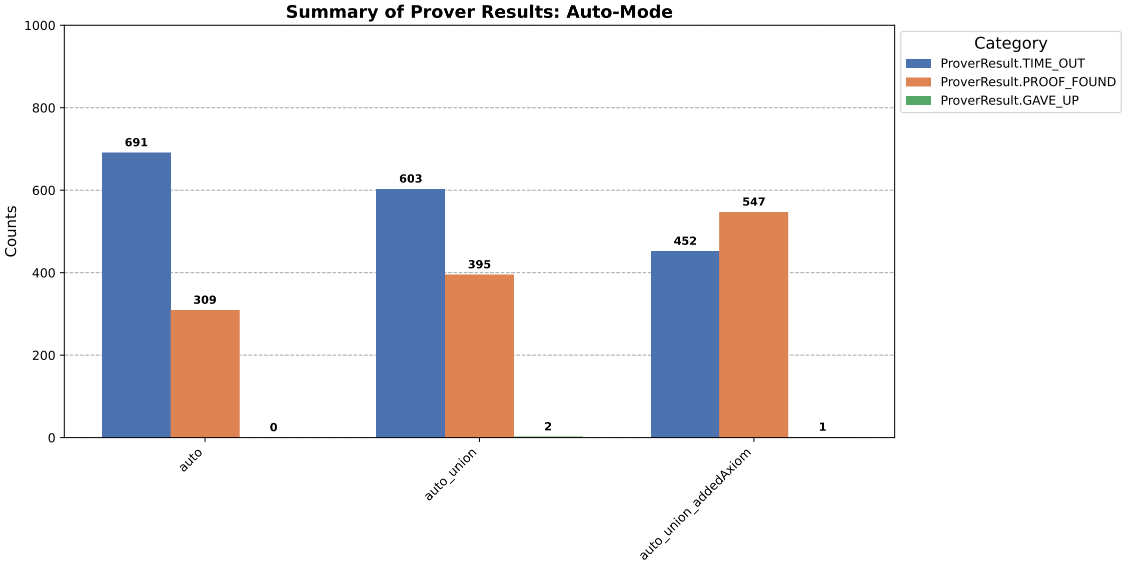

Figure 4.12: Summary of prover results in Auto mode

5

Conclusion

Conclusion.

6

Future work

Future work.

Bibliography

Alvez, J., Hermo, M., Lucio, P., and Rigau, G. (May 2017). Automatic White-Box Testing of First-Order Logic Ontologies. Journal of Logic and Computation *29*. DOI: 10.1093/logcom/exz001.

Alvez, J., Lucio, P., and Rigau, G. (Oct. 2014). Adimen-SUMO: Reengineering an Ontology for First-Order Reasoning. International Journal on Semantic Web and Information Systems *8*, 80–116. DOI: 10.4018/jswis.2012100105.

Bayerkuhnlein, M. (2023). SUMO Wrestling Winograd Schemas. In Proceedings of a University Research Project, Bamberg, Germany. URL: https://www.uni-bamberg.de.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics *5*, 135–146. DOI: 10.1162/TACL_A_00051. URL: https://doi.org/10.1162/tacl_a_00051.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

Furbach, U., Kramer, T., and Schon, C. (2019). Names Are Not Just Sound and Smoke: Word Embeddings for Axiom Selection. In Automated Deduction – CADE 27, (Fontaine, P., ed.). Cham: Springer International Publishing, pp. 250–268. ISBN: 978-3-030-29436-6. DOI: 10.1007/978-3-030-29436-6_15.

Hoder, K. and Voronkov, A. (2011). Sine Qua Non for Large Theory Reasoning. In Automated Deduction – CADE-23, (Bjorner, N. and Sofronie-Stokkermans, V., eds.). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 299–314. ISBN: 978-3-642-22438-6.

Jakobs, O. and Schon, C. (2024). Context-Specific Selection of Commonsense Knowledge Using Large Language Models. In KI 2024: Advances in Artificial Intelligence, (Hotho, A. and Rudolph, S., eds.). Cham: Springer Nature Switzerland, pp. 218–231. ISBN: 978-3-031-70893-0.

Johnson-Laird, P.N. (1989). Mental Models. Foundations of Cognitive Science, (Posner, M.I., ed.). The MIT Press, pp. 469–499.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-Thought Vectors. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015 (NeurIPS), (Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., and Garnett, R., eds.). Montreal, Quebec, Canada, pp. 3294–3302.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning,

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Advances in Neural Information Processing Systems 26 (NeurIPS), (Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q., eds.), pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, G.A. and Charles, W.G. (1991). Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6, 1–28. DOI: 10.1080/01690969108406936. eprint: https://doi.org/10.1080/01690969108406936. URL: https://doi.org/10.1080/01690969108406936.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2001), ACM, pp. 2–9. DOI: 10.1145/505168.505170. URL: https://doi.org/10.1145/505168.505170.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (Moschitti, A., Pang, B., and Daelemans, W., eds.). Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Reimers, N. and Gurevych, I. (Nov. 2019a). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. URL: https://arxiv.org/abs/1908.10084.

Reimers, N. and Gurevych, I. (2019b). *Sentence-Transformers: Multilingual Sentence and Image Embeddings using BERT and Co.* URL: https://github.com/UKPLab/sentence-transformers.

Schon, C. (2023). Associative Reasoning for Commonsense Knowledge. In KI 2023: Advances in Artificial Intelligence, (Seipel, D. and Steen, A., eds.). Cham: Springer Nature Switzerland, pp. 170–183. ISBN: 978-3-031-42608-7.

Schulz, S., Cruanes, S., and Vukmirovic, P. (2019). Faster, Higher, Stronger: E 2.3. In Proceedings of the 27th International Conference on Automated Deduction (CADE-27), (Fontaine, P., ed.). Lecture Notes in Artificial Intelligence (LNAI) 11716. Springer, pp. 495–507.