

Conversational Neuro-Symbolic Commonsense Reasoning

Forough Arabshahi

Carnegie Mellon University
farabsha@cs.cmu.edu

Jennifer Lee

Carnegie Mellon University
jennife4@cs.cmu.edu

Mikayla Gawarecki

Carnegie Mellon University
mgawarec@andrew.cmu.edu

Kathryn Mazaitis

Carnegie Mellon University
Pittsburgh, PA 15213
krivard@cs.cmu.edu

Amos Azaria

Ariel University
amos.azaria@ariel.ac.il

Tom Mitchell

Carnegie Mellon University
tom.mitchell@cs.cmu.edu

Abstract

One aspect of human commonsense reasoning is the ability to make *presumptions* about daily experiences, activities and social interactions with others. We propose a new commonsense reasoning benchmark where the task is to uncover commonsense *presumptions* implied by imprecisely stated natural language commands in the form of if-then-because statements. For example, in the command “If it snows at night then wake me up early because I don’t want to be late for work” the speaker relies on commonsense reasoning of the listener to infer the implicit presumption that it must snow enough to cause traffic slowdowns. Such if-then-because commands are particularly important when users instruct conversational agents. We release a benchmark data set for this task, collected from humans and annotated with commonsense presumptions. We develop a neuro-symbolic theorem prover that extracts multi-hop reasoning chains and apply it to this problem. We further develop an interactive conversational framework that evokes commonsense knowledge from humans for completing reasoning chains.

1 Introduction

Despite the remarkable success of artificial intelligence (AI) and machine learning in the last few decades, commonsense reasoning remains an unsolved problem at the heart of AI [1–3]. Commonsense allows us humans to engage in conversations with one another and to convey our thoughts efficiently, without the need to specify much detail [4]. For example, if Alice asks Bob to “wake her up early whenever it snows at night” so that she can get to work on time, Alice assumes that Bob will wake her up only if it snows enough to cause traffic slowdowns, and only if it is a working day. Alice does not explicitly state these conditions since Bob makes such *presumptions* without much effort thanks to his common sense. A study, in which we collected if-then commands from human subjects, revealed that humans often under-specify conditions in their statements; perhaps because they are used to speaking with other humans who possess the common sense needed to infer their more specific intent by making presumptions about their statement. The inability to make these presumptions is one of the main reasons why it is challenging for computers to engage in natural sounding conversations with humans.

In an attempt to enable this, we propose a new commonsense reasoning benchmark where the task is to infer *commonsense presumptions* in commands of the form “If *state* holds Then perform *action* Because I want to achieve *goal*.” The reason for including the “because” clause in the commands is that some presumptions are ambiguous without knowing the user’s purpose, or goal. For instance, if Alice’s goal in the above example was to see snow for the first time, Bob would have presumed that even a snow flurry would be excuse enough to wake her up. Since humans frequently omit details when stating such commands, a computer possessing common sense should be able

Table 1: Statistics of if-then-because commands collected from a pool of human subjects. The table shows four distinct types of *because*-clauses we found, the count of commands of each type, examples of each and their corresponding commonsense presumption annotations. *Restricted domain* includes commands whose state is limited to checking email, calendar, maps, alarms, and weather. *Everyday domain* includes commands concerning more general day-to-day activities. Annotations are tuples of (index, presumption) where index shows the starting word index of where the missing presumption should be in the command - highlighted with a red arrow. Index starts at 0 and is calculated for the original command.

Domain	if clause	then clause	because clause	Example	Annotation: Commonsense Presumptions	Count
Restricted domain	state	action	goal	If it's going to rain in the afternoon ⁽¹⁾ then remind me to bring an umbrella ⁽²⁾ because I want to remain dry	(8, and I am outside) (15, before I leave the house)	76
Restricted domain	state	action	anti-goal	If I have an upcoming bill payment ⁽¹⁾ then remind me to pay it ⁽²⁾ because I don't want to pay a late fee	(7, in the next few days) (13, before the bill payment deadline)	3
Restricted domain	state	action	modifier	If my flight ⁽¹⁾ is from 2am to 4am then book me a supershuttle ⁽²⁾ because it will be difficult to find ubers.	(3, take off time) (13, for 2 hours before my flight take off time)	2
Restricted domain	state	action	conjunction	If I receive emails about sales on basketball shoes ⁽¹⁾ then let me know ⁽²⁾ because I need them and I want to save money.	(9, my size) (13, there is a sale)	2
SUM						83
Everyday domain	state	action	goal	If there is an upcoming election ⁽¹⁾ then remind me to register ⁽²⁾ and vote ⁽³⁾ because I want my voice to be heard.	(6, in the next few months) (6, and I am eligible to vote) (11, to vote), (13, in the election)	55
Everyday domain	state	action	anti-goal	If it's been two weeks since my last call with my mentee and I don't have an upcoming appointment with her ⁽¹⁾ then remind me to send her an email ⁽²⁾ because we forgot to schedule our next chat	(21, in the next few days) (29, to schedule our next appointment)	4
Everyday domain	state	action	modifier	If I have difficulty sleeping ⁽¹⁾ then play a lullaby because it soothes me.	(5, at night)	12
Everyday domain	state	action	conjunction	If the power goes out ⁽¹⁾ then when it comes back on remind me to restart the house furnace because it doesn't come back on by itself and I want to stay warm	(5, in the Winter)	6
SUM						77

to infer the hidden *presumptions*; that is, the additional unstated conditions on the If and/or Then portion of the command (refer to Tab. 1 to see some examples).

We propose an approach that infers such missing presumptions, by extracting a chain of reasoning that shows how the commanded action will achieve the desired goal when the state holds. Whenever any additional reasoning steps appear in this reasoning chain, they are output by our system as assumed implicit presumptions associated with the command. For our reasoning method we propose a neuro-symbolic interactive, conversational approach, in which the computer combines its own common sense knowledge with conversationally evoked knowledge provided by a human user. The reasoning chain is extracted using our neuro-symbolic theorem prover that learns sub-symbolic representations (embeddings) for logical statements, making it robust to variations of natural language encountered in a conversational interaction setting.

Contributions We have three main contributions. 1) We propose a benchmark task for commonsense reasoning and release a data set containing if-then-because commands, annotated with commonsense presumptions. 2) We present a system called CORGI (COMmonsense ReasoninG by Instruction) that performs soft logical inference. We propose a neuro-symbolic theorem prover and apply it to extract a multi-hop reasoning chain that reveals commonsense presumptions. 3) We equip CORGI with a conversational interaction mechanism that enables it to collect just-in-time commonsense knowledge from humans. Our user-study shows (a) the plausibility of relying on humans to evoke commonsense knowledge and (b) the effectiveness of our theorem prover, enabling us to extract reasoning chains for up to 45% of the studied tasks¹.

Related Work The literature on commonsense reasoning dates back to the very beginning of the field of AI [2, 5, 6] and is studied in several contexts. One aspect focuses on building a large knowledge base (KB) of commonsense facts. Projects like CYC [7], ConceptNet [8–10] and ATOMIC [11, 12] are examples of such KBs (see [2] for a comprehensive list). Recently, [13] proposed COMET, a generative model trained on ConceptNet and ATOMIC, that generates commonsense facts. These KBs provide background knowledge for tasks requiring common sense. However, it is known that knowledge bases are incomplete, and most have ambiguities and inconsistencies [2] that must be clarified for particular reasoning tasks. Therefore, we argue that reasoning engines can benefit greatly from a conversational interaction strategy to ask humans about their missing or

¹Our code and data are publicly available here <https://github.com/ForoughA/CORGI>

inconsistent knowledge. Closest in nature to this proposal is the work by Hixon et al., [14] on relation extraction through conversation for question answering. The advent of intelligent agents and advancements in natural language processing have given learning from conversational interactions a good momentum in the last few years [15–23].

A current challenge in commonsense reasoning is lack of benchmarks [2]. Benchmark tasks in commonsense reasoning include the Winograd Schema Challenge (WSC) [1], its variations [24], and its recently scaled up counterpart, Winogrande [3]; ROCStories [25], COPA [26], and \mathcal{ART} [27], where the task is to choose a plausible outcome, cause or explanation for an input scenario. Most of these benchmarks have a multiple choice design format. However, in the real world the computer is usually not given multiple choice questions. None of these benchmarks targets the extraction of unspoken details in a natural language statement, which is a challenging task for computers known since the 1970’s [4].

CORGI has a neuro-symbolic logic theorem prover. Neuro-symbolic systems are hybrid models that leverage the robustness of connectionist methods and the soundness of symbolic reasoning to effectively integrate learning and reasoning [28, 29]. They have shown promise in different areas of logical reasoning ranging from classical logic to propositional logic, probabilistic logic, abductive logic, and inductive logic [30–35]. To the best of our knowledge, neuro-symbolic solutions for commonsense reasoning have not been proposed before. Examples of commonsense reasoning engines are: AnalogySpace [36, 37] that uses dimensionality reduction, and [6] that uses the event calculus formal language. TensorLog [38] converts a first-order logical database into a factor graph and proposes a differentiable strategy for belief propagation over the graph. DeepProbLog [31] developed a probabilistic logic programming language that is suitable for applications containing categorical variables. Contrary to our approach, both these methods do not learn embeddings for logical rules that are needed to make CORGI robust to natural language variations. Therefore, we propose an end-to-end differentiable solution that uses a Prolog [39] proof trace to learn rule embeddings from data. Our proposal is closest to the neural programmer interpreter [40] that uses the trace of algorithms such as addition and sort to learn their execution. The use of Prolog for performing multi-hop logical reasoning has been studied in [41, 42]. These methods perform Inductive Logic Programming to learn rules from data, and are not applicable to our problem. DeepLogic [43] and [44, 45] also learn representations for logical rules using neural networks. Very recently, transformers were used for temporal logic [46] and to do multi-hop reasoning [47] using logical facts and rules stated in natural language. A purely connectionist approach to reasoning suffers from some limitations. For example, the input token size limit of transformers restricts [47] to small knowledge bases. Moreover, generalizing to arbitrary number of variables or an arbitrary inference depth is not trivial for them. Since symbolic reasoning can inherently handle all these challenges, a hybrid approach to reasoning takes the burden of handling them off of the neural component.

2 Proposed Commonsense Reasoning Benchmark

The benchmark task that we propose in this work is that of uncovering hidden *commonsense presumptions* given an input if-then-because command. Formally, the commands follow the general format “if $\langle \text{state holds} \rangle$ then $\langle \text{perform action} \rangle$ because $\langle \text{I want to achieve goal} \rangle$ ”. We refer to the *if*-clause as the *state*, the *then*-clause as the *action* and the *because*-clause as the *goal*. These natural language commands were collected from a pool of human subjects (more details in the Appendix). The data is annotated with unspoken commonsense presumptions by a team of annotators. Tab. 1 shows the statistics of the data and annotated examples from the data. We collected two sets of if-then-because commands. The first set contains 83 commands targeted at a *state* that can be observed by a computer/mobile phone (which is checking emails, calendar, maps, alarms, and weather). The second set contains 77 commands whose *state* is about day-to-day events and activities. 81% of the commands over both sets qualify as “if $\langle \text{state} \rangle$ then $\langle \text{action} \rangle$ because $\langle \text{goal} \rangle$ ”. The remaining 19% differ in the categorization of the *because*-clause (see Tab. 1); common alternate clause types included anti-goals (“...because I don’t want to be late”), modifications of the state or action (“... because it will be difficult to find an Uber”), or conjunctions including at least one non-goal type. Note that we did not instruct the subjects to give us data from these categories, rather we uncovered them after data collection. Note that commonsense benchmarks such as the Winograd Schema Challenge [1] included a similar number of examples (100) when first introduced [24].

Lastly, after collecting the data we discovered that the if-then-because commands given by humans can be categorized into several different logic templates. The discovered logic templates are given in Table 5 in the Appendix. Our neuro-symbolic theorem prover uses a general reasoning strategy

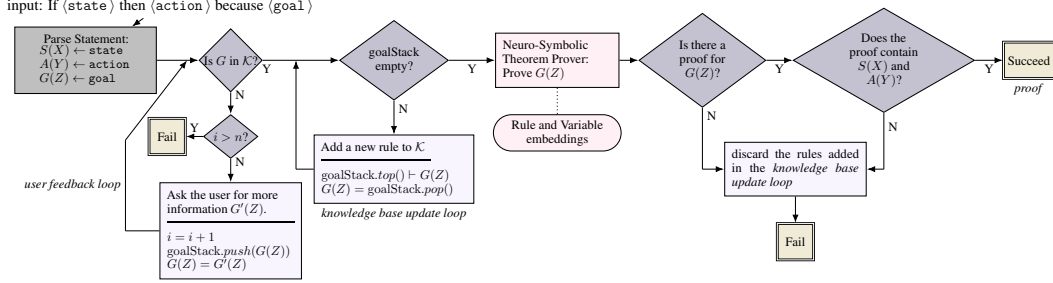


Figure 1: CORGI’s flowchart. The input is an if-then-because command e.g., “if it snows tonight then wake me up early because I want to get to work on time”. The input is parsed into its logical form representation (for the prev example, $S(X) = \text{weather}(\text{snow}, \text{Precipitation})$). If CORGI succeeds, it outputs a proof tree for the because-clause or goal (parsed into $G(Z) = \text{get}(i, \text{work}, \text{on_time})$). The output proof tree contains commonsense presumptions for the input statement (Fig 2 shows an example). If the predicate G does not exist in the knowledge base, \mathcal{K} , (Is G in \mathcal{K} ?), we have missing knowledge and cannot find a proof. Therefore, we extract it from a human in the *user feedback loop*. At the heart of CORGI is a neuro-symbolic theorem prover that learns rule and variable embeddings to perform a proof (Alg.1). *goalStack* and the loop variable i are initialized to empty and 0 respectively, and $n = 3$. *italic text* in the figure represents descriptions that are referred to in the main text.

that can address all reasoning templates. However, in an extended discussion in the Appendix, we explain how a reasoning system, including ours, could potentially benefit from these logic templates.

3 Method

Background and notation The system’s commonsense knowledge is a KB, denoted \mathcal{K} , programmed in a Prolog-like syntax. We have developed a modified version of Prolog, which has been augmented to support several special features (types, soft-matched predicates and atoms, etc). Prolog [39] is a declarative logic programming language that consists of a set of predicates whose arguments are atoms, variables or predicates. A predicate is defined by a set of rules (Head $:-$ Body.) and facts (Head.), where Head is a predicate, Body is a conjunction of predicates, and $:-$ is logical implication. We use the notation $S(X)$, $A(Y)$ and $G(Z)$ to represent the logical form of the *state*, *action* and *goal*, respectively where S , A and G are predicate names and X , Y and Z indicate the *list* of arguments of each predicate. For example, for *goal* = “I want to get to work on time”, we have $G(Z) = \text{get}(i, \text{work}, \text{on_time})$. Prolog can be used to logically “prove” a query (e.g., to prove $G(Z)$) using the backward chaining algorithm (see the Appendix - Prolog Background).

3.1 CORGI: Commonsense Reasoning by Instruction

CORGI takes as input a natural language command of the form “if $\langle \text{state} \rangle$ then $\langle \text{action} \rangle$ because $\langle \text{goal} \rangle$ ” and infers commonsense presumptions by extracting a chain of commonsense knowledge that explains how the commanded *action* achieves the *goal* when the *state* holds. For example from a high level, for the command in Fig. 2 CORGI outputs (1) if it snows more than two inches, then there will be traffic, (2) if there is traffic, then my commute time to work increases, (3) if my commute time to work increases then I need to leave the house earlier to ensure I get to work on time (4) if I wake up earlier then I will leave the house earlier. Formally, this reasoning chain is a proof tree (proof trace) shown in Fig.2. As shown, the proof tree includes the commonsense presumptions.

CORGI’s architecture is depicted in Figure 1. In the first step, the if-then-because command goes through a parser that extracts the *state*, *action* and *goal* from it and converts them to their logical form representations $S(X)$, $A(Y)$ and $G(Z)$, respectively. For example, the *action* “wake me up early” is converted to $\text{wake}(\text{me}, \text{early})$. The parser is presented in the Appendix (Sec. Parsing).

The proof trace is obtained by finding a proof for $G(Z)$, using \mathcal{K} and the context of the input if-then-because command. In other words, $S(X) \cap A(Y) \cap \mathcal{K} \rightarrow G(Z)$. One challenge is that even the largest knowledge bases gathered to date are incomplete, making it virtually infeasible to prove an arbitrary input $G(Z)$. Therefore, CORGI is equipped with a conversational interaction strategy which enables it to prove a query by combining its own commonsense knowledge with

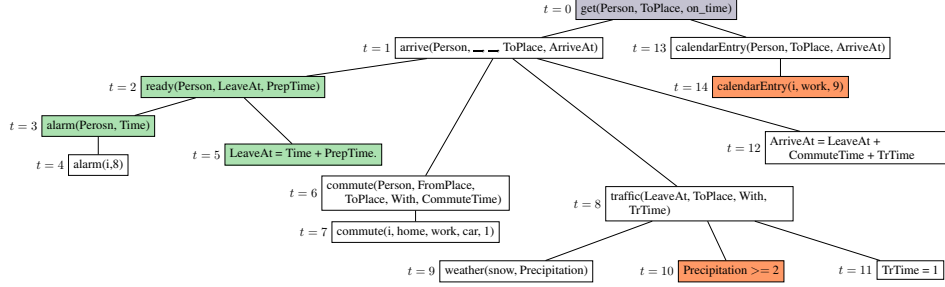


Figure 2: Sample proof tree for the *because*-clause of the statement: “If it snows tonight then wake me up early because I want to get to work on time”. Proof traversal is depth-first from left to right (t gives the order). Each node in the tree indicates a rule’s head, and its children indicate the rule’s body. For example, the nodes highlighted in green indicate the rule $\text{ready}(\text{Person}, \text{LeaveAt}, \text{PrepTime}) \text{ :- } \text{alarm}(\text{Person}, \text{Time}) \wedge \text{LeaveAt} = \text{Time} + \text{PrepTime}$. The goal we want to prove, $G(Z) = \text{get}(\text{Person}, \text{ToPlace}, \text{on_time})$, is in the tree’s root. If a proof is successful, the variables in $G(Z)$ get grounded (here Person and ToPlace are grounded to i and work , respectively). The highlighted orange nodes are the uncovered commonsense presumptions.

conversationally evoked knowledge provided by a human user in response to a question from CORGI (*user feedback loop* in Fig.1). There are 4 possible scenarios that could occur when designing such a conversational knowledge extraction strategy.

- Ⓐ The user understands the question, but does not know the answer.
- Ⓑ The user misunderstands the question and responds with an undesired answer.
- Ⓒ The user understands the question and provides a correct answer, but the system fails to understand the user due to:
 - Ⓒ.1 limitations of natural language understanding.
 - Ⓒ.2 variations in natural language, which result in misalignment of the data schema in the knowledge base and the data schema in the user’s mind.
- Ⓓ The user understands the questions and provides the correct answer and the system successfully parses and understands it.

CORGI’s different components are designed such that they address the above challenges, as explained below. Since our benchmark data set deals with day-to-day activities, it is unlikely for scenario Ⓐ to occur. If the task required more specific domain knowledge, Ⓐ could have been addressed by choosing a pool of domain experts. Scenario Ⓑ is addressed by asking informative questions from users. Scenario Ⓒ.1 is addressed by trying to extract small chunks of knowledge from the users piece-by-piece. Specifically, the choice of what to ask the user in the *user feedback loop* is deterministically computed from the user’s goal. The first step is to ask how to achieve the user’s stated goal, and CORGI expects an answer that gives a sub-goal. In the next step, CORGI asks how to achieve the sub-goal the user just mentioned. The reason for this piece-by-piece knowledge extraction is to ensure that the language understanding component can correctly parse the user’s response. CORGI then adds the extracted knowledge from the user to \mathcal{K} in the *knowledge update loop* shown in Fig.1. Missing knowledge outside this goal/sub-goal path is not handled, although it is an interesting future direction. Moreover, the model is user specific and the knowledge extracted from different users are not shared among them. Sharing knowledge raises interesting privacy issues and requires handling personalized conflicts and falls out of the scope of our current study.

Scenario Ⓒ.2, caused by the variations of natural language, results in semantically similar statements to get mapped into different logical forms, which is unwanted. For example, “make sure I am awake early morning” vs. “wake me up early morning” will be parsed into different logical forms $\text{awake}(i, \text{early_morning})$ and $\text{wake}(\text{me}, \text{early_morning})$, respectively although they are semantically similar. This mismatch prevents a logical proof from succeeding since the proof strategy relies on exact match in the unification operation (see Appendix). This is addressed by our neuro-symbolic theorem prover (Fig.1) that learns vector representations (embeddings) for logical rules and variables and uses them to perform a logical proof through soft unification. If the theorem prover can prove the user’s goal, $G(Z)$, CORGI outputs the proof trace (Fig.2) returned by its theorem prover and succeeds. In the next section, we explain our theorem prover in detail. We revisit scenarios Ⓐ – Ⓓ in detail in the discussion section and show real examples from our user study.

4 Neuro-Symbolic Theorem Proving

Our Neuro-Symbolic theorem prover is a neural modification of *backward chaining* and uses the vector similarity between rule and variable embeddings for unification. In order to learn these embeddings, our theorem prover learns a general proving strategy by training on proof traces of successful proofs. From a high level, for a given query our model maximizes the probability of choosing the correct rule to pick in each step of the backward chaining algorithm. This proposal is an adaptation of Reed et al.’s Neural Programmer-Interpreter [40] that learns to execute algorithms such as addition and sort, by training on their execution trace.

In what follows, we represent scalars with lowercase letters, vectors with bold lowercase letters and matrices with bold uppercase letters. $\mathbf{M}^{rule} \in \mathbb{R}^{n_1 \times m_1}$ denotes the embedding matrix for the rules and facts, where n_1 is the number of rules and facts and m_1 is the embedding dimension. $\mathbf{M}^{var} \in \mathbb{R}^{n_2 \times m_2}$ denotes the variable embedding matrix, where n_2 is the number of all the atoms and variables in the knowledge base and m_2 is the variable embedding dimension. Our knowledge base is type coerced, therefore the variable names are associated with their types (e.g., `alarm(Person, Time)`)

Learning The model’s core consists of an LSTM network whose hidden state indicates the next rule in the proof trace and a proof termination probability, given a query as input. The model has a feed forward network that makes variable binding decisions. The model’s training is fully supervised by the proof trace of a query given in a depth-first-traversal order from left to right (Fig. 2). The trace is sequentially input to the model in the traversal order as explained in what follows. In step $t \in [0, T]$ of the proof, the model’s input is $\epsilon_t^{inp} = (\mathbf{q}_t, \mathbf{r}_t, (\mathbf{v}_t^1, \dots, \mathbf{v}_t^\ell))$ and T is the total number of proof steps. \mathbf{q}_t is the query’s embedding and is computed by feeding the predicate name of the query into a character RNN. \mathbf{r}_t is the concatenated embeddings of the rules in the parent and the left sister nodes in the proof trace, looked up from \mathbf{M}^{rule} . For example in Fig.2, \mathbf{q}_3 represents the node at proof step $t = 3$, \mathbf{r}_3 represents the rule highlighted in green (parent rule), and \mathbf{r}_4 represents the fact `alarm(i, 8)`. The reason for including the left sister node in \mathbf{r}_t is that the proof is conducted in a left-to-right depth first order. Therefore, the decision of what next rule to choose in each node is dependent on both the left sisters and the parent (e.g. the parent and the left sisters of the node at step $t = 8$ in Fig. 2 are the rules at nodes $t = 1$, $t = 2$, and $t = 6$, respectively). The arguments of the query are presented in $(\mathbf{v}_t^1, \dots, \mathbf{v}_t^\ell)$ where ℓ is the arity of the query predicate. For example, \mathbf{v}_3^1 in Fig 2 is the embedding of the variable `Person`. Each \mathbf{v}_t^i for $i \in [0, \ell]$, is looked up from the embedding matrix \mathbf{M}^{var} . The output of the model in step t is $\epsilon_t^{out} = (c_t, \mathbf{r}_{t+1}, (\mathbf{v}_{t+1}^1, \dots, \mathbf{v}_{t+1}^\ell))$ and is computed through the following equations

$$\mathbf{s}_t = f_{enc}(\mathbf{q}_t), \quad \mathbf{h}_t = f_{lstm}(\mathbf{s}_t, \mathbf{r}_t, \mathbf{h}_{t-1}), \quad (1)$$

$$c_t = f_{end}(\mathbf{h}_t), \quad \mathbf{r}_{t+1} = f_{rule}(\mathbf{h}_t), \quad \mathbf{v}_{t+1}^i = f_{var}(\mathbf{v}_t^i), \quad (2)$$

where \mathbf{v}_{t+1}^i is a probability vector over all the variables and atoms for the i^{th} argument, \mathbf{r}_{t+1} is a probability vector over all the rules and facts and c_t is a scalar probability of terminating the proof at step t . f_{enc} , f_{end} , f_{rule} and f_{var} are feed forward networks with two fully connected layers, and f_{lstm} is an LSTM network. The trainable parameters of the model are the parameters of the feed forward neural networks, the LSTM network, the character RNN that embeds \mathbf{q}_t and the rule and variable embedding matrices \mathbf{M}^{rule} and \mathbf{M}^{var} .

Our model is trained end-to-end. In order to train the model parameters and the embeddings, we maximize the log likelihood probability given below

$$\theta^* = \underset{\theta}{argmax} \sum_{\epsilon^{out}, \epsilon^{in}} \log(P(\epsilon^{out} | \epsilon^{in}; \theta)), \quad (3)$$

where the summation is over all the proof traces in the training set and θ is the trainable parameters of the model. We have

$$\log(P(\epsilon^{out} | \epsilon^{in}; \theta)) = \sum_{t=1}^T \log(P(\epsilon_t^{out} | \epsilon_1^{in} \dots \epsilon_{t-1}^{in}; \theta)) \quad (4)$$

$$\begin{aligned} \log P(\epsilon_t^{out} | \epsilon_1^{in} \dots \epsilon_{t-1}^{in}; \theta) &= \log P(\epsilon_t^{out} | \epsilon_{t-1}^{in}; \theta) \\ &= \log P(c_t | \mathbf{h}_t) + \log P(\mathbf{r}_{t+1} | \mathbf{h}_t) + \log \sum_i P(\mathbf{v}_{t+1}^i | \mathbf{v}_t^i). \end{aligned} \quad (5)$$

Table 2: percentage of successful reasoning tasks for different user types. In no-feedback, user responses are not considered in the proof attempt. In soft unification CORGI uses our proposed neuro-symbolic theorem prover. In the Oracle scenario, the theorem prover has access to oracle embeddings and soft unification is 100% accurate.

CORGI variations	Novice User	Expert User
No-feedback	0%	0%
Soft unification	15.61%	35.00%
Oracle unification	21.62%	45.71%

Table 3: Sample dialogs of 2 novice users in our study. CORGI’s responses are noted in italics.

Successful task
If it’s going to rain in the afternoon then remind me to bring an umbrella because I want to remain dry. <i>How do I know if “I remain dry”?</i>
If I have my umbrella. <i>How do I know if “I have my umbrella”?</i>
If you remind me to bring an umbrella. <i>Okay, I will perform “remind me to bring an umbrella” in order to achieve “I remain dry”.</i>
Failed task
If it’s going to rain in the afternoon then remind me to bring an umbrella because I want to remain dry. <i>How do I know if “I remain dry”?</i>
If I have my umbrella. <i>How do I know if “I have my umbrella”?</i>
If it’s in my office. <i>How do I know if “it’s in my office”?</i>
...

Where the probabilities in Equation (5) are given in Equations (2). The inference algorithm for proving is given in the Appendix, section Inference.

5 Experiment Design

The knowledge base, \mathcal{K} , used for all experiments is a small handcrafted set of commonsense knowledge. See Tab.6 in the Appendix for examples. \mathcal{K} includes general information about time, restricted-domains such as setting alarms and notifications, emails, and so on, as well as commonsense knowledge about day-to-day activities. \mathcal{K} contains a total of 228 facts and rules. Among these, there are 189 everyday-domain and 39 restricted domain facts and rules. We observed that most of the if-then-because commands require everyday-domain knowledge for reasoning, even if they are restricted-domain commands (see Table 3 for example).

Our Neuro-Symbolic theorem prover is trained on proof traces collected by proving automatically generated query’s to \mathcal{K} using sPyrolog². M^{rule} and M^{var} are initialized randomly and with GloVe embeddings [48], respectively, where $m_1 = 256$ and $m_2 = 300$. Since \mathcal{K} is type-coerced (e.g. Time, Location, . . .), initializing the variables with pre-trained word embeddings helps capture their semantics and improves the performance. The neural components of the theorem prover is implemented in PyTorch [49] and the prover is built on top of sPyrolog.

User Study In order to assess CORGI’s performance, we ran a user study. We selected 10 goal-type if-then-because commands from the dataset in Table 1 and used each as the prompt for a reasoning task. We had 28 participants in the study, 4 of which were experts closely familiar with CORGI and its capabilities. The rest were undergraduate and graduate students with the majority being in engineering or computer science fields and some that majored in business administration or psychology. These users had never interacted with CORGI prior to the study (novice users). Each person was issued the 10 reasoning tasks, taking on average 20 minutes to complete all 10.

Solving a reasoning task consists of participating in a dialog with CORGI as the system attempts to complete a proof for the goal of the current task; see sample dialogs in Tab. 3. The task succeeds if CORGI is able to use the answers provided by the participant to construct a reasoning chain (proof) leading from the goal to the state and action. We collected 469 dialogues in our study.

The user study was run with the architecture shown in Fig. 1. We used the participant responses from the study to run a few more experiments. We (1) Replace our theorem prover with an *oracle* prover that selects the optimal rule at each proof step in Alg. 1 and (2) attempt to prove the goal without using any participant responses (*no-feedback*). Tab. 2 shows the success rate in each setting.

Discussion In this section, we analyze the results from the study and provide examples of the 4 scenarios in Section 3.1 that we encountered. As hypothesized there, scenario \mathfrak{A} hardly occurred. We did encounter scenario \mathfrak{B} , however. The study’s dialogs show that some users provided means of *sensing* the goal rather than the *cause* of the goal. For example, for the reasoning task “*If there are thunderstorms in the forecast within a few hours then remind me to close the windows because I want to keep my home dry*”, in response to the system’s prompt “*How do I know if ‘I keep my*

²<https://github.com/leonweber/spyrolog>

home dry’?” a user responded *“if the floor is not wet”* as opposed to an answer such as *“if the windows are closed”*. Moreover, some users did not pay attention to the context of the reasoning task. For example, another user responded to the above prompt (same reasoning task) with *“if the temperature is above 80”*! Overall, we noticed that CORGI’s ability to successfully reason about an if-then-because statement was heavily dependent on whether the user knew how to give the system what it needed, and not necessarily what it asked for; see Table 3 for an example. As it can be seen in Table 2, expert users are able to more effectively provide answers that complete CORGI’s reasoning chain, likely because they know that regardless of what CORGI asks, the object of the dialog is to connect the because goal back to the knowledge base in some series of if-then rules (goal/sub-goal path in Sec.3.1). Therefore, one interesting future direction is to develop a dynamic context-dependent Natural Language Generation method for asking more effective questions.

We would like to emphasize that although it seems to us, humans, that the previous example requires very simple background knowledge that likely exists in SOTA large commonsense knowledge graphs such as ConceptNet³, ATOMIC⁴ or COMET [13], this is not the case (verifiable by querying them online). For example, for queries such as *“the windows are closed”*, COMET-ConceptNet generative model⁵ returns knowledge about blocking the sun, and COMET-ATOMIC generative model⁶ returns knowledge about keeping the house warm or avoiding to get hot; which while being correct, is not applicable in this context. For *“my home is dry”*, both COMET-ConceptNet and COMET-ATOMIC generative models return knowledge about house cleaning or house comfort. On the other hand, the fact that 40% of the novice users in our study were able to help CORGI reason about this example with responses such as *“If I close the windows”* to CORGI’s prompt, is an interesting result. This tells us that conversational interactions with humans could pave the way for commonsense reasoning and enable computers to extract just-in-time commonsense knowledge, which would likely either not exist in large knowledge bases or be irrelevant in the context of the particular reasoning task. Lastly, we re-iterate that as conversational agents (such as Siri and Alexa) enter people’s lives, leveraging conversational interactions for learning has become a more realistic opportunity than ever before.

In order to address scenario C.1, the conversational prompts of CORGI ask for specific small pieces of knowledge that can be easily parsed into a predicate and a set of arguments. However, some users in our study tried to provide additional details, which challenged CORGI’s natural language understanding. For example, for the reasoning task *“If I receive an email about water shut off then remind me about it a day before because I want to make sure I have access to water when I need it.”*, in response to the system’s prompt *“How do I know if ‘I have access to water when I need it.’?”* one user responded *“If I am reminded about a water shut off I can fill bottles”*. This is a successful knowledge transfer. However, the parser expected this to be broken down into two steps. If this user responded to the prompt with *“If I fill bottles”* first, CORGI would have asked *“How do I know if ‘I fill bottles’?”* and if the user then responded *“if I am reminded about a water shut off”* CORGI would have succeeded. The success from such conversational interactions are not reflected in the overall performance mainly due to the limitations of natural language understanding.

Table 2 evaluates the effectiveness of conversational interactions for proving compared to the no-feedback model. The 0% success rate there reflects the incompleteness of \mathcal{K} . The improvement in task success rate between the no-feedback case and the other rows indicates that when it is possible for users to contribute useful common-sense knowledge to the system, performance improves. The users contributed a total number of 96 rules to our knowledge base, 31 of which were unique rules. Scenario C.2 occurs when there is variation in the user’s natural language statement and is addressed with our neuro-symbolic theorem prover. Rows 2-3 in Table 2 evaluate our theorem prover (*soft unification*). Having access to the optimal rule for unification does still better, but the task success rate is not 100%, mainly due to the limitations of natural language understanding explained earlier.

6 Conclusions

In this paper, we introduced a benchmark task for commonsense reasoning that aims at uncovering unspoken intents that humans can easily uncover in a given statement by making presumptions supported by their common sense. In order to solve this task, we propose CORGI (COMmon-sense

³<http://conceptnet.io/>

⁴https://mosaickg.apps.allenai.org/kg_atomic

⁵https://mosaickg.apps.allenai.org/comet_conceptnet

⁶https://mosaickg.apps.allenai.org/comet_atomic

ReasoninG by Instruction) which is a neuro-symbolic theorem prover and performs commonsense reasoning by initiating a conversation with a user. CORGI has access to a small knowledge base of commonsense facts and completes it through time as she interacts with the user. We further conduct a user study that indicates the possibility of using conversational interactions with humans for evoking commonsense knowledge and verifies the effectiveness of our proposed theorem prover.

References

- [1] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.
- [2] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.
- [3] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. 2019.
- [4] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.
- [5] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.
- [6] Erik T Mueller. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann, 2014.
- [7] Douglas B Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Sheperd. Cyc: toward programs with common sense. *Communications of the ACM*, 33(8):30–49, 1990.
- [8] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.
- [9] Catherine Havasi, Robyn Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer, 2007.
- [10] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017.
- [11] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. *arXiv preprint arXiv:1811.00146*, 2018.
- [12] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*, 2018.
- [13] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*, 2019.
- [14] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 851–861, 2015.
- [15] Amos Azaria, Jayant Krishnamurthy, and Tom M Mitchell. Instructable intelligent personal agent. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [16] Igor Labutov, Shashank Srivastava, and Tom Mitchell. Lia: A natural language programmable personal assistant. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 145–150, 2018.
- [17] Shashank Srivastava. *Teaching Machines to Classify from Natural Language Interactions*. PhD thesis, Samsung Electronics, 2018.
- [18] Dan Goldwasser and Dan Roth. Learning from natural instructions. *Machine learning*, 94(2):205–232, 2014.
- [19] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 729–738, 2019.

- [20] Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Advances in Neural Information Processing Systems*, pages 2942–2951, 2018.
- [21] Toby Jia-Jun Li, Igor Labutov, Xiaohan Nancy Li, Xiaoyi Zhang, Wenze Shi, Wanling Ding, Tom M Mitchell, and Brad A Myers. Appinite: A multi-modal interface for specifying data descriptions in programming by demonstration using natural language instructions. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 105–114. IEEE, 2018.
- [22] Toby Jia-Jun Li, Yuanchun Li, Fanglin Chen, and Brad A Myers. Programming iot devices by demonstration using mobile apps. In *International Symposium on End User Development*, pages 3–17. Springer, 2017.
- [23] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. Sugilite: creating multimodal smartphone automation by demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6038–6049. ACM, 2017.
- [24] Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*, 2020.
- [25] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, 2017.
- [26] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.
- [27] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.
- [28] Artur d’Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiák, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*, 2015.
- [29] Tarek R Besold, Artur d’Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017.
- [30] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. 2019.
- [31] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, pages 3749–3759, 2018.
- [32] Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. Neural logic machines. In *International Conference on Learning Representations (ICLR)*, 2019.
- [33] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori. Integrating learning and reasoning with deep logic models. *arXiv preprint arXiv:1901.04195*, 2019.
- [34] Zhi-Hua Zhou. Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):76101, 2019.
- [35] Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.
- [36] Robyn Speer, Catherine Havasi, and Henry Lieberman. Analogyspace: Reducing the dimensionality of common sense knowledge. In *AAAI*, volume 8, pages 548–553, 2008.
- [37] Catherine Havasi, Robyn Speer, James Pustejovsky, and Henry Lieberman. Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent systems*, 24(4):24–35, 2009.

- [38] William W Cohen. Tensorlog: A differentiable deductive database. *arXiv preprint arXiv:1605.06523*, 2016.
- [39] Alain Colmerauer. An introduction to prolog iii. In *Computational Logic*, pages 37–79. Springer, 1990.
- [40] Scott Reed and Nando De Freitas. Neural programmer-interpreters. *arXiv preprint arXiv:1511.06279*, 2015.
- [41] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems*, pages 3788–3800, 2017.
- [42] L Weber, P Minervini, J Münchmeyer, U Leser, and T Rocktäschel. Nlprolog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, Volume 1: Long Papers*, volume 57. ACL (Association for Computational Linguistics), 2019.
- [43] Nuri Cingillioglu and Alessandra Russo. Deeplogic: Towards end-to-end differentiable logical reasoning. *arXiv preprint arXiv:1805.07433*, 2018.
- [44] Tim Rocktäschel, Matko Bošnjak, Sameer Singh, and Sebastian Riedel. Low-dimensional embeddings of logic. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 45–49, 2014.
- [45] William Yang Wang and William W Cohen. Learning first-order logic embeddings via matrix factorization. In *IJCAI*, pages 2132–2138, 2016.
- [46] Bernd Finkbeiner, Christopher Hahn, Markus N Rabe, and Frederik Schmitt. Teaching temporal logics to neural networks. *arXiv preprint arXiv:2003.04218*, 2020.
- [47] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020.
- [48] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [50] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 2017.
- [51] Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246, 2017.

Appendix

Data Collection

Data collection was done in two stages. In the first stage, we collected if-then-because commands from human subjects. In the second stage, a team of annotators annotated the data with common-sense presumptions. Below we explain the details of the data collection and annotation process.

In the data collection stage, we asked a pool of human subjects to write commands that follow the general format: if \langle state holds \rangle then \langle perform action \rangle because \langle i want to achieve goal \rangle . The subjects were given the following instructions at the time of data collection:

“Imagine the two following scenarios:

Scenario 1: Imagine you had a personal assistant that has access to your email, calendar, alarm, weather and navigation apps, what are the tasks you would like the assistant to perform for your day-to-day life? And why?

Scenario 2: Now imagine you have an assistant/friend that can understand anything. What would you like that assistant/friend to do for you?

Our goal is to collect data in the format “If then because”

After the data was collected, a team of annotators annotated the commands with additional presumptions that the human subjects have left unspoken. These presumptions were either in the *if*-clause and/or the *then*-clause and examples of them are shown in Tables 1 and 4

Table 4: Example if-then-because commands in the data and their annotations. Annotations are tuples of (index, missing text) where index shows the starting word index of where the missing text should be in the command. Index starts at 0 and is calculated for the original utterance.

Utterance	Annotation
If the temperature (↓) is above 30 degrees (↓) then remind me to put the leftovers from last night into the fridge because I want the leftovers to stay fresh	(2, inside) (7, Celsius)
If it snows (↓) tonight (↓) then wake me up early because I want to arrive to work early	(3, more than two inches) (4, and it is a working day)
If it’s going to rain in the afternoon (↓) then remind me to bring an umbrella (↓) because I want to stay dry	(8, when I am outside) (15, before I leave the house)

Logic Templates

As explained in the main text, we uncovered 5 different logic templates, that reflect humans’ reasoning, from the data after data collection. The templates are listed in Table 5. In what follows, we will explain each template in detail using the examples of each template listed in Tab. 5.

In the blue template (Template 1), the `state` results in a “bad state” that causes the *not* of the goal. The speaker asks for the `action` in order to avoid the bad state and achieve the goal. For instance, consider the example for the blue template in Table 5. The `state` of snowing a lot at night, will result in a bad state of traffic slowdowns which in turn causes the speaker to be late for work. In order to overcome this bad state. The speaker would like to take the `action`, waking up earlier, to account for the possible slowdowns cause by snow and get to work on time.

In the orange template (Template 2), performing the `action` when the `state` holds allows the speaker to achieve the goal and not performing the `action` when the `state` holds prevents the speaker from achieving the goal. For instance, in the example for the orange template in Table 5

Table 5: Different reasoning templates of the statements that we uncovered, presumably reflecting how humans logically reason. \wedge , \neg , $:-$ indicate logical and, negation, and implication, respectively. $action_h$ is an action that is hidden in the main utterance and $action(state)$ indicates performing the $action$ when the $state$ holds.

Logic template	Example	Count
1. $(\neg(goal) :- state) \wedge$ $(goal :- action(state))$	If it snows tonight then wake me up early because I want to arrive to work on time	65
2. $(goal :- action(state)) \wedge$ $(\neg(goal) :- \neg(action(state)))$	If I am walking to a meeting then remind me who else is there because I want to be prepared for the meeting	50
3. $(goal :- action_h) \wedge$ $(action_h :- action(state))$	If we are approaching Fall then remind me to buy flower bulbs because I want to make sure I have a pretty Spring garden.	17
4. $(goal :- state) \wedge$ $(\neg(goal) :- action(state))$	If I am at the grocery store but I have a trip coming up in the next week then remind me not to buy perishables because they will go bad while I am away	5
5. other	If tomorrow is a holiday then ask me if I want to disable or change my alarms because I don't want to wake up early if I don't need to go to work early.	23

the speaker would like to know who the attendees of a meeting are when the speaker is walking to that meeting so that the speaker is prepared for the meeting and that if the speaker is not reminded of this, he/she will not be able to properly prepare for the meeting.

In the green template (Template 3), performing the $action$ when the $state$ holds allows the speaker to take a hidden $action$ that enables him/her to achieve the desired $goal$. For example, if the speaker is reminded to buy flower bulbs close to the Fall season, he/she will buy and plant the flowers (hidden $action$ s) that allows the speaker to have a pretty spring garden.

In the purple template (Template 4), the $goal$ that the speaker has stated is actually a goal that they want to *avoid*. In this case, the $state$ causes the speaker's $goal$, but the speaker would like to take the $action$ when the $state$ holds to achieve the opposite of the $goal$. For the example in Tab. 1, if the speaker has a trip coming up and he/she buys perishables the perishables would go bad. In order for this not to happen, the speaker would like to be reminded not to buy perishables to avoid them going bad while he/she is away.

The rest of the statements are categorized under the "other" category. The majority of these statements contain conjunction in their $state$ and are a mix of the above templates. A reasoning engine could potentially benefit from these logic templates when performing reasoning. We provide more detail about this in the Extended Discussion section in the Appendix.

Prolog Background

Prolog [39] is a declarative logic programming language. A Prolog program consists of a set of predicates. A predicate has a name (functor) and $N \geq 0$ arguments. N is referred to as the arity of the predicate. A predicate with functor name F and arity N is represented as $F(T_1, \dots, T_N)$ where T_i 's, for $i \in [1, N]$, are the arguments that are arbitrary Prolog terms. A Prolog term is either an atom, a variable or a compound term (a predicate with arguments). A variable starts with a capital letter (e.g., Time) and atoms start with small letters (e.g. monday). A predicate defines a relationship between its arguments. For example, `isBefore(monday, tuesday)` indicates that the relationship between Monday and Tuesday is that, the former is before the latter.

A predicate is defined by a set of clauses. A clause is either a Prolog *fact* or a Prolog *rule*. A Prolog rule is denoted with `Head :- Body.`, where the Head is a predicate, the Body is a conjunction (\wedge) of predicates, $:-$ is logical implication, and period indicates the end of the clause. The previous rule is an if-then statement that reads "if the Body holds then the Head holds". A fact is a rule whose body always holds, and is indicated by `Head.`, which is equivalent to `Head :- true`. Rows 1-4 in Table 6 are rules and rows 5-8 are facts.

Prolog can be used to logically "prove" whether a specific query holds or not (For example, to prove that `isAfter(wednesday, thursday)?` is false or that `status(i, dry, tuesday)?` is true using the

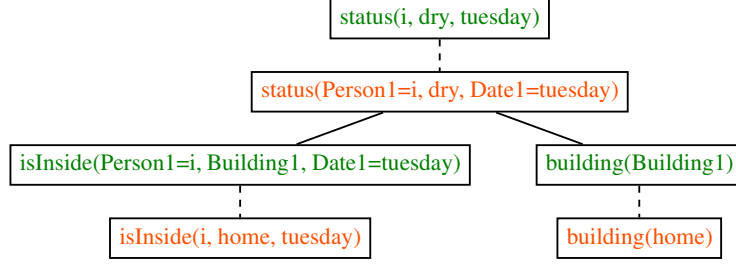


Figure 3: Sample simplified proof tree for query `status(i, dry, tuesday)`. dashed edges show successful unification, orange nodes show the head of the rule or fact that is retrieved by the unification operator in each step and green nodes show the query in each proof step. This proof tree is obtained using the Prolog program or \mathcal{K} shown in Tab. 6. In the first step, unification goes through all the rules and facts in the table and retrieves rule number 2 whose head unifies with the query. This is because the query and the rule head’s functor name is `status` and they both have 3 arguments. Moreover, the arguments all match since `Person1` grounds to atom `i`, grounded atom `dry` matches in both and variable `Date1` grounds to `tuesday`. In the next step, the proof iterates through the predicates in the rule’s body, which are `isInside(i, Building1, tuesday)` and `building(Building1)`, to recursively prove them one by one using the same strategy. Each of the predicates in the body become the new query to prove and proof succeeds if all the predicates in the body are proved. Note that once the variables are grounded in the head of the rule they are also grounded in the rule’s body.

Program in Table 6). The proof is performed through *backward chaining*, which is a backtracking algorithm that usually employs a depth-first search strategy implemented recursively. In each step of the recursion, the input is a query (goal) to prove and the output is the proof’s success/failure. In order to prove a query, a rule or fact whose head *unifies* with the query is retrieved from the Prolog program. The proof continues recursively for each predicate in the body of the retrieved rule and succeeds if all the statements in the body of a rule are *true*. The base case (leaf) is when a fact is retrieved from the program.

At the heart of backward chaining is the *unification* operator, which matches the query with a rule’s head. Unification first checks if the functor of the query is the same as the functor of the rule head. If they are the same, unification checks the arguments. If the number of arguments or the arity of the predicates do not match unification fails. Otherwise it iterates through the arguments. For each argument pair, if both are grounded atoms unification succeeds if they are exactly the same grounded atoms. If one is a variable and the other is a grounded atom, unification grounds the variable to the atom and succeeds. If both are variables unification succeeds without any variable grounding. The backwards chaining algorithm and the unification operator is depicted in Figure 3.

Parsing

The goal of our parser is to extract the `state`, `action` and `goal` from the input utterance and convert them to their logical forms $S(X)$, $A(Y)$, and $G(Z)$, respectively. The parser is built using Spacy [50]. We implement a relation extraction method that uses Spacy’s built-in dependency parser. The language model that we used is the `en_coref_lg-3.0.0` released by Hugging face⁷. The predicate name is typically the sentence verb or the sentence root. The predicate’s arguments are the subject, objects, named entities and noun chunks extracted by Spacy. The output of the relation extractor is matched against the knowledge base through rule-based mechanisms including string matching to decide whether the parsed logical form exists in the knowledge base. If a match is found, the parser re-orders the arguments to match the order of the arguments of the predicate retrieved from the knowledge base. This re-ordering is done through a type coercion method. In order to do type coercion, we use the types released by Allen AI in the Aristo tuple KB v1.03 Mar 2017 Release [51] and have added more entries to it to cover more nouns. The released types file is a dictionary that maps different nouns to their types. For example, `doctor` is of type `person` and `Tuesday` is of type

⁷https://github.com/huggingface/neuralcoref-models/releases/download/en_coref_lg-3.0.0/en_coref_lg-3.0.0.tar.gz

Algorithm 1 Neuro-Symbolic Theorem Prover

Input: goal $G(Z)$, M^{rule} , M^{var} , Model parameters, threshold T_1, T_2, k
Output: Proof P

$r_0 \leftarrow \mathbf{0}$ $\triangleright \mathbf{0}$ is a vector of 0s
 $P = \text{PROVE}(G(Z), r_0, [])$
function $\text{PROVE}(Q, r_t, \text{stack})$
 embed Q using the character RNN to obtain q_t
 input q_t and r_t to the model and compute r_{t+1} (Equation (2))
 compute c_t (Equation (2))
 $\{R^1 \dots R^k\} \leftarrow \text{From } M^{rule} \text{ retrieve } k \text{ entries corresponding to the top } k \text{ entries of } r_{t+1}$
 for $i \in [0, k]$ **do**
 $SU \leftarrow \text{SOFT_UNIFY}(Q, \text{head}(R^i))$
 if $SU == \text{False}$ **then**
 continue to $i + 1$
 else
 if $c_t > T_2$ **then**
 return stack
 add R^i to stack
 $\text{PROVE}(\text{Body}(R^i), r_{t+1}, \text{stack})$ \triangleright Prove the body of R^i
 return stack
function $\text{SOFT_UNIFY}(G, H)$
 if $\text{arity}(G) \neq \text{arity}(H)$ **then**
 return False
 Use M^{var} to compute cosine similarity S_i for all corresponding variable pairs in G and H
 if $S_i > T_1 \quad \forall \quad i \in [0, \text{arity}(G)]$ **then**
 return True
 else
 return False

date. If no match is found, the parsed predicate will be kept as is and CORGI tries to evoke relevant rules conversationally from humans in the *user feedback loop* in Figure 1.

We would like to note that we refrained from using a grammar parser, particularly because we want to enable open-domain discussions with the users and save the time required for them to learn the system’s language. As a result, the system will learn to adapt to the user’s language over time since the background knowledge will be accumulated through user interactions, therefore it will be adapted to that user. A negative effect, however, is that if the parser makes a mistake, error will propagate onto the system’s future knowledge. This is an interesting future direction that we are planning to address.

Inference

The inference algorithm for our proposed neuro-symbolic theorem prover is given in Alg. 1. In each step t of the proof, given a query Q , we calculate q_t and r_t from the trained model to compute r_{t+1} . Next, we choose k entries of M^{rule} corresponding to the top k entries of r_{t+1} as candidates for the next proof trace. k is set to 5 and is a tuning parameter. For each rule in the top k rules, we attempt to do variable/argument unification by computing the cosine similarity between the arguments of Q and the arguments of the rule’s head. If all the corresponding pair of arguments in Q and the rule’s head have a similarity higher than threshold, $T_1 = 0.9$, unification succeeds, otherwise it fails. If unification succeeds, we move to prove the body of that rule. If not, we move to the next rule.

Extended Discussion

Table 7 shows the performance breakdown with respect to the logic templates in Table 5. Currently, CORGI uses a general theorem prover that can prove all the templates. The large variation in performance indicates that taking into account the different templates would improve the performance. For example, the low performance on the green template is expected, since CORGI currently does not support the extraction of a hidden action from the user, and interactions only support extraction

Table 6: Examples of the commonsense rules and facts in \mathcal{K}

1	<code>isEarlierThan(Time1,Time2) :- isBefore(Time1,Time3), isEarlierThan(Time3,Time2).</code>
2	<code>status(Person1, dry, Date1) :- isInside(Person1, Building1, Date1), building(Building1).</code>
3	<code>status(Person1, dry, Date1) :- weatherBad(Date1, _), carry(Person1, umbrella, Date1), isOutside(Person1, Date1).</code>
4	<code>notify(Person1, corgi, Action1) :- email(Person1, Action1).</code>
5	<code>isBefore(monday, tuesday).</code>
6	<code>has(house, window).</code>
7	<code>isInside(i, home, tuesday).</code>
8	<code>building(home).</code>

of missing goals. This interesting observation indicates that, even within the same benchmark, we might need to develop several reasoning strategies to solve reasoning problems. Therefore, even if CORGI adapts a general theorem prover, accounting for logic templates in the conversational knowledge extraction component would allow it to achieve better performance on other templates.

Table 7: Number of successful reasoning tasks vs number of attempts under different scenarios. In CORGI’s Oracle unification, soft unification is 100% accurate. LT stands for Logic Template and LTi refers to template i in Table 5.

CORGI	LT1	LT2	LT3	LT5
Oracle Unification	24%	38%	11%	0%