A Computational Model of Tractable Reasoning — taking inspiration from cognition*

Lokendra Shastri
Department of Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
(215) 898-2661; shastri@cis.upenn.edu

Abstract

Polynomial time complexity is the usual 'threshold' for distinguishing the tractable from the intractable and it may seem reasonable to adopt this notion of tractability in the context of knowledge representation and reasoning. It is argued that doing so may be inappropriate in the context of common sense reasoning underlying language understanding. A more stringent criteria of tractability is proposed. A result about reasoning that is tractable in this stronger sense is outlined. Some unusual properties of tractable reasoning emerge when the formal specification is grounded in a neurally plausible architecture.

1 Introduction

Understanding language is a complex task. It involves among other things, carrying out inferences in order to establish referential and causal coherence, generate expectations, and make predictions. Nevertheless we can understand language at the rate of *several hundred words per minute* [Carpenter and Just, 1977]. This rapid rate of language understanding suggests that we can (and do) perform a wide range of inferences very rapidly, automatically and without conscious effort — as though they are a *reflex* response of our cognitive apparatus. In view of this such reasoning may be described as *reflexive* [Shastri, 1991].

As an example of reflexive reasoning consider the sentence 'John seems to have suicidal tendencies, he has joined the Columbian drug enforcement agency.' We can understand this sentence spontaneously and without any deliberate effort even though, doing so involves the use of background knowledge and reasoning. Informally, this reasoning may be as follows: joining the Columbian drug enforcement agency has dangerous consequences, and as John may be aware of this, his decision to join the agency suggests that he has suicidal tendencies. As another example of reflexive reasoning consider the inference 'John owns a car' upon hearing 'John bought a Rolls-Royce'. We can make this inference effortlessly even though it requires multiple steps of inference using background knowledge such as Rolls-Royce is a car and if *x* buys *y* then *x* owns *y.*

Not all reasoning is, and as complexity theory tells us, cannot be, reflexive. We contrast reflexive reasoning with *reflective* reasoning — reasoning that requires reflection, conscious deliberation, and at times, the use of external props such as paper and pencil (e.g., solving logic puzzles, doing cryptarithmetic, or planning a vacation).

2 Reflexive reasoning necessitates a strong notion of tractability

In order to quantify the notion of reflexive reasoning introduced above, let us make a few observations about such reasoning.

- *Reflexive reasoning occurs with respect to a large body of background knowledge.* A serious attempt at compiling common sense knowledge suggests that our background knowledge base may contain as many as 10^7 to 10^8 items [Guha and Lenat, 1990]. This should not be very surprising given that this knowledge includes, besides other things, our knowledge of naive physics and naive psychology; facts about ourselves, our family, friends, colleagues, history and geography; our knowledge of artifacts, sports, art, music; some basic principles of science and mathematics; and our models of social, civic, and political interactions.
- Items in the background knowledge base are fairly stable and persist for a long-time once they are acquired. Hence this knowledge is best described as *long-term* knowledge and we will refer to this body of knowledge as the long-term knowledge base (LTKB).
- Episodes of reflexive reasoning are triggered by 'small' inputs. In the context of language understanding, an input (typically) corresponds to a sentence that would map into a small number of assertions. For example, the input 'John bought a Rolls Royce' maps into just one assertion (or a few, depending on the underlying representation). The critical observation is that *the size of the input, \In\,*

*This work was supported by NSF grants IRI 88-05465 and the ARO grant DAAL 03-89-C-0031.

is insignificant compared to the size of the long-term knowledge base \LTKB\[1] [2]

- The vast difference in the magnitude of \LTKB\ (about 10^8) and \In\ (a few) becomes crucial when analyzing the tractability of common sense reasoning. Given the actual values of \In\ that occur during common sense reasoning, there is a distinct possibility that the overall cost of a derivation may be dominated by the "fixed" contribution of \LTKBI. Thus we cannot ignore the cost attributable to \LTKB\ and we must analyze the complexity of reasoning in terms of\LTKB\ as well as \In\.

In view of the magnitude of \LTKB\, even a cursory analysis suggests that any inference procedure whose time complexity is quadratic or worse in \LTKB\ cannot provide a plausible computational account of reflexive reasoning. However, a process that is polynomial in \In\ remains viable.

2.1 Time complexity of reflexive reasoning

Observe that although the size of a person's \LTKB\ increases considerably from, say, age seven to thirty, the time taken by a person to understand natural language does not. This suggests that the time taken by an episode of reflexive reasoning does not depend on the \LTKB\. In view of this it is proposed that a realistic criteria of tractability for reflexive reasoning is one where the time taken by an episode of reflexive reasoning is independent of \LTKB\ and only depends on the depth of the derivation tree associated with the inference.[3]

2.2 Space complexity of reflexive reasoning

The expected size of the LTKB also rules out any computational scheme whose space requirement is quadratic (or higher) in the size of the KB. For example, the brain has only about 10^{12} cells most of which are involved in processing of sensorimotor information. Hence even a linear space requirement is fairly generous and leaves room only for a modest 'constant of proportionality'. In view of this, it is proposed that the admissible space requirement of a model of reflexive reasoning be no more than linear in \LTKB\.

To summarize, it is proposed that as far as (reflexive) reasoning underlying language understanding is concerned, the appropriate notion of tractability is one where

- the reasoning time is independent of \LTKB\ and is only dependent on \In\ and the depth of the derivation tree associated with the inference, and
- the associated space requirement, i.e., the space required to encode the LTKB plus the space required to hold the working memory during reasoning should be no worse than linear in \LTKB\.

In spite of the apparent significance of reflexive reasoning there have been very few attempts at developing a computational account of such inference. Some past exceptions being Fahlman's work on NETL [1979] and Shastri's work on a connectionist semantic memory [1988]. However these models dealt primarily with inheritance and classification within an IS~A hierarchy. Holldobler [I 990] and Ullman and van Gelder [1988] nave proposed parallel systems for performing quite complex logical inferences, however, both these systems have unrealistic space requirements. The number of nodes in Holldobler's system is quadratic in the the size of the knowledge base (KB) the number of processors required by Ullman and van Gelder is even higher. Ullman and van Gelder treat the number of nodes required to encode the background KB as a fixed cost, and hence, do not refer to its size in computing the space complexity of their system. If the size of such a KB is taken into account, the number of processors required by their system turns out to be a high degree polynomial.

A significant amount of work has been done by researchers in knowledge representation and reasoning to identify classes of limited inference that can be performed efficiently (e.g., see [Frisch and Allen, 1982]; [Brachman and Levesque, 1984]; [Patel-Schneider, 1985]; [Dowling and Gallier, 1984]; [Levesque, 1988]; [Selman and Levesque, 1989]; [Mc A Hester, 1990]; [Bylander et a/., 1991]; [Kautz and Selman, 1991]). This work has covered a wide band of the complexity spectrum but none that meets the strong tractability requirement discussed above. Most results stipulate polynomial time complexity, restrict inference in implausible ways (e.g., by excluding chaining of rules), and/or deal with limited expressiveness (e.g., deal only with propositions).

3 A tractable reasoning class

Below we describe a class of reasoning that is tractable with reference to the criteria stated above. The characterization of such a class is different (but analogous) for forward and backward reasoning. In this paper we will focus on backward reasoning.

Some definitions:

Let us define *rules* to be first-order sentences of the form:

$$\forall x_1, ..., x_m\, [P_1(...) \wedge P_2(...) ... \wedge P_n(...) \Rightarrow \exists z_1, ... z_l\, Q(...)]$$

where the arguments of P_i's are elements of $\{x_1, ... x_m\}$, and an argument of Q is either an element of $\{x_1, ... x_m\}$, an element of $\{z_1, ... z_l\}$, or a constant. □

Any variable that occurs in multiple argument positions in the antecedent of a rule is a *pivotal* variable. □

[1] A small input may, however, lead to a potentially large number of elaborate inferences. For example, the input 'John bought a Rolls-Royce' may generate a number of reflexive inferences such as 'John bought a car', 'John owns a car', 'John has a driver's license', 'John is perhaps a wealthy man', etc.

[2] Some of these inferences may be 'soft' inferences, but the issue of deductive versus evidential nature of inferences is irrelevant to our current concerns.

[3] The restriction that the reasoning time be independent of \LTKB\ may seem overly strong and one might argue that perhaps logarithmic time may be acceptable. Our belief that the stronger notion of effectiveness is relevant, however, is borne out by results which demonstrate that there does exists a class of reasoning that can be performed in time independent of \LTKB\.

Note that the notion of a pivotal variable is local to a rule.

A rule is *balanced* if all pivotal variables occurring in the rule also appear in its consequent. □

For example, the rule $\forall x, y, z\ P(x,y) \wedge R(x,z) \Rightarrow S(y,z)$ is not balanced because the pivotal variable x does not occur in the consequent. On the other hand, the rule $\forall x, y, z\ P(x,y) \wedge R(x,z) \Rightarrow S(x,z)$ is balanced because the pivotal variable x does occur in the consequent. The fact that y does not appear in the consequent is immaterial because y occurs only once in the antecedent and hence, is not a pivotal variable.

Facts are partial or complete instantiations of predicates. Thus facts are atomic formulae of the form $P(t_1, t_2 ... t_k)$ where *t^a* are either constants or distinct existentially quantified variables. □

Queries have the same form as facts. Let us distinguish between *yes-no* queries and wh-queries. A query, all of whose arguments are bound to constants corresponds to the *yes-no* query: 'Does the query follow from the rules and facts encoded in the long-term memory of the system?* A query with existentially quantified variables, however, has several interpretations. For example, the query P(a, x), where a is a constant and *x* is an existentially quantified argument, may be viewed as the *yes-no* query: 'Does P(a,x) follow from the rules and facts for some value of x?' Alternately this query may be viewed as the wh-query: Tor what values of *x* does *P(a,x)* follow from the rules and facts in the system's long-term memory?' D

Consider a query Q and a LTKB consisting of facts and balanced rules. A derivation of Q obtained by backward chaining is *threaded* if all pivotal variables occurring in the derivation get bound and their bindings can be traced back to the bindings introduced in *Q*. □

Given a LTKB consisting of facts and balanced rules, a *reflexive* query is one for which there exists a threaded proof. □

3.1 A class of tractable reasoning

The worst-case time for answering a reflexive *yes-no* query, *Q,* is proportional to $V|In|^V$ where:

- $|In|$ is the number of *distinct* constants in *Q.*
- *V* is as follows: Let V_i be the arity of the predicate P_i. Then *V* equals $max(V_i)$, i ranging over all the predicates in the LTKB.
- *d* equals the depth of the shallowest derivation of Q given the LTKB.

Observe that the worst-case time is i) *independent* of *\LTKB\,* ii) polynomial in *\In* and iii) only proportional to *d.*

As observed in Section 2, while *\LTKB* may be as much as 10^8, *\In* is simply the number of (distinct) 'entities' referred to in the input. In the context of natural language understanding, *\In* would be quite small (typically, less than 5). We also expect V, the maximum arity of predicates in the LTKB to be quite small.

An answer to a *wh-query* can also be computed in time proportional to $V|In|^V$d, except that *\In* now equals the arity of the query predicate Q.

The *space* requirement is *linear* in *\LTKB* and polynomial in *\In\.* This includes the cost of encoding the LTKB as well as the cost of maintaining the dynamic state of the 'working memory' during reasoning.

An informal explanation of the result

The number of times a predicate *P* may get instantiated in a threaded derivation of a query cannot exceed $|In|^V$. This follows from the observation that *P* has at most *V* arguments and each of these can get bound to at most *\In* distinct constants. Since each predicate instantiation can contain at most *V* bindings, the propagation of argument bindings from one predicate to another can be carried out in time proportional to $V|In|^V$. This assumes that the correspondence (specified by the rules in the LTKB) between the arguments of the antecedent and consequent predicates are hard-wired.

It can be shown that the propagation of argument bindings from multiple predicates to a predicate can be carried out in parallel (see [Mani and Shastri, 1992] for a possible implementation of such a parallel binding propagation scheme). This means that the time required to carry out one step of a parallel breadth-first derivation is only proportional to $V|In|^V$. It follows that the time required to carry out a *d* step parallel derivation is proportional to $V|In|^V d$.

Lower bound nature of above result

In general, derivations that involve unbalanced rules or those that do not satisfy the *threaded* property cannot be computed in time independent of *\LTKB\,* if the available space is no more than *linear* in *\LTKB* [Dietz *et al,* 1993). This result follows from the observations that i) the *common-element* problem, i.e., the problem of determining whether two sets share a common element, can be reduced to the problem of computing a derivation involving unbalanced rules and/or non-threaded derivations, ii) the *sorting* problem can be reduced to the common-element problem, and iii) the lower bound on the sorting problem is *nlogn* (where n would correspond to *\LTKB\).* Thus derivations involving unbalanced rules and non-threaded derivations may not be computed in time independent of *\LTKB* unless one makes use of more than *linear* space.

3.2 Worst-case versus expected case

The above result offers a worst-case characterization which assumes that during the derivation, *all variables will get instantiated with all possible bindings involving constants in Q.* This will not be the case in a typical situation. In fact it may be conjectured that in a typical episode of reasoning, the actual time will seldom exceed 50d (see next section).

4 A neurally motivated model of tractable reasoning

We have proposed a neurally plausible model (SHRUTI) that can encode a LTKB of the type described above, together with a term hierarchy and perform a class of forward as well as backward reasoning with extreme efficiency [Shastri and Ajjanagadde, 1990]; [Ajianagadde and Shastri, 1991]; [Mani and Shastri, 1991]; [Mani and Shastri, 1992]; [Shastri, 1992]. SHRUTI can draw inferences in time that is only proportional to the *depth* of the shallowest derivation leading to the conclusion. A SHRUTI like model has also been used by Henderson [1992] to design a parser for English. The parser's speed is independent of the size of the lexicon and the grammar, and it offers a natural explanation for a variety of data on long distance dependencies and center embedding.

If we set aside SHRUTUs ability to perform terminological reasoning, the class of reasoning that SHRUTI can perform efficiently is a subclass of the class of reasoning specified in the previous section. The additional restrictions placed on SHRUTI's reasoning ability are motivated by gross constraints on the speed at which humans can perform reflexive reasoning and gross neurophysiologies parameters such as:

1. π_{max}, the maximum period at which nodes can be expected to sustain synchronous activity,
2. w, the tolerance or the minimum lead/lag that must be allowed between the spiking of two nodes that are firing in synchrony,
3. the time it takes a cluster of synchronous nodes to drive a connected cluster of nodes to fire in synchrony.

The details of the model are beyond the scope of this paper and the reader is referred to [Shastri and Ajjanagadde, 1990]. Let us however, state the additional constraints on the class of reasoning SHRUTI can perform.

4.1 Additional constraints on the reasoning performed by SHRUTI

SHRUTI can encode a LTKB of facts and *balanced* rules and answer yes to any *reflexive yes-no* query in time proportional to the *depth* of the shallowest derivation leading to a derivation of the query provided:

1. The number of distinct constants specified in the query does not exceed k_1, where *k* is bounded by $\pi_{max}/\omega)$ (biological data suggests that k_1 is small, perhaps between 5 and 10).

   The model suggests that as long as the number of entities introduced by the query is 5 or less, there will essentially be no cross-talk among the facts inferred during reasoning. If more than 5 entities occur, the window of synchrony would have to shrink appropriately in order to accommodate all the entities. As this window shrinks, the possibility of cross-talk between bindings would increase until eventually, the cross-talk would become excessive and disrupt the system's ability to perform systematic reasoning. The biological data suggests that a neurally plausible *upper bound* on the number of distinct entities that can occur in the reasoning process is about 10. Of course, these entities may occur in multiple facts and participate in a number of inferences.

   It may be significant that the bound on the number of entities that may be referenced by the active facts during a derivation relates well to 7 ± 2, the robust measure of short-term memory capacity [Miller, 1956]. Note however, that SHRUTI does not place a small limit on the number of *facts* that can be simultaneously active — indeed a very large number of facts can be involved in each derivation carried out by SHRUTI.

2. During the processing of the query, each predicate may only be instantiated at most k_2 times.

   Note that this restriction only applies to run-time or 'dynamic' instantiations of predicates and not to iong-term' facts stored in the system. As argued in [Shastri, 1992] a plausible values of k_2 is somewhere between 3-5. Also, k_2 need not be the same for all predicates. The application of a SHRUTI-like model to parsing by Henderson also suggests that a value of k_2 under 3 may be sufficient for parsing English sentences.

Some typical retrieval and inference timings

If we set system parameters of SHRUTI to some neurally motivated values, SHRUTI demonstrates that a system made up of simple and slow neuron-like elements can perform a wide range of inferences (both forward, backward and those involving a type hierarchy) within a few hundred milliseconds.

If we choose the period of oscillation of nodes to be 20 milliseconds, assume that nodes can synchronize within two periods of oscillations and pick k_2 equal to 3, SHRUTI takes 320 milliseconds to infer 'John is jealous of Tom' after being given the dynamic facts 'John loves Susan' and 'Susan loves Tom' (this involves the rule 'if x loves *y* and *y* loves *z* then x is jealous of *z).* The system takes 260 milliseconds to infer 'John is a sibling of Jack' given 'Jack is a sibling of John' (this involves the rule 'if x is a sibling of *y* then *y* is a sibling of x). Similarly, the system takes 320 milliseconds to infer 'Susan *owns* a car' after its internal state is initialized to represent 'Susan *bought* a Rolls-Royce' (using the rule 'if x buys *y* then x owns y' and the *IS-A* relation, 'Rolls-Royce is a car').

If SHRUTI's long-term memory contains the fact 'John bought a Rolls-Royce', SHRUTI takes 140 milliseconds, 420 milliseconds, and 740 milliseconds, respectively, to answer 'yes' to the queries 'Did John buy a Rolls-Royce', 'Does John own a car?' and 'Can John sell a car?' (the last query also makes use of the rule 'if x owns *y* then x can sell *y).* Note that the second and third queries also involve inferences using rules as well as *IS-A* relations.

The above times are independent of *\LTKB* and do not increase when additional rules, facts, and *IS-A* relationships are added. If anything, these times may decrease if a new rule is added that leads to a shorter inference path.

5 Conclusion

We have proposed a criteria for tractable reasoning that is appropriate in the context of common sense reasoning underlying language understanding. We have suggested that an appropriate measure of tractability for such reasoning is one where the time complexity is independent of, and the space complexity is no more than linear in, the size of the long-term knowledge base. We have also identified a class of reasoning that is tractable in this sense. This characterization of tractability can be further refined by cognitive and biological considerations. This work suggests that the expressiveness and the inferential ability of a representation and reasoning systems may be limited in unusual ways to arrive at extremely efficient yet fairly powerful knowledge representation and reasoning systems.

References

[Ajjanagadde and Shastri, 199l] V.G. Ajjanagadde and L. Shastri. Rules and variables in neural nets. Neural Computation, 3:121-134.

[Brachman and Levesque, 1984] R. Brachman and H. Levesque. The tractability of Subsumption in frame-based description languages. In Proceedings of AAAI-84, the fourth national conference on artificial intelligence. Morgan Kaufman.

[Bylander et al., 1991] T. Bylander, D. Allemang, M. C. Tanner, J. R. Josephson. The computational complexity of abduction. Artificial Intelligence, 47(1-3), 25-60.

[Carpenter and Just, 1977] P.A. Carpenter and M.A. Just. Reading Comprehension as Eyes See It. In: Cognitive Processes in Comprehension.

[Dietz etai, 1993] P. Dietz, D. Krizanc, S. Rajasekaran, L. Shastri. A lower-bound result for the common element problem and its implication for reflexive reasoning. Forthcoming Technical Report, Department of Computer and Information Science, Univ. of Pennsylvania.

[Dowling and Gallier, 1984] W.F. Dowling and J.H. Gallier. Linear time algorithm for testing the satisfiability of propositional horn formula. Journal of Logic Programming, 3:267-284.

[Fahlman, 1979] S.E. Fahlman. NETL: A system for representing real-world knowledge, MIT Press.

[Frisch and Allen, 1982] A.M. Frisch and J.F. Allen. Knowledge retrieval as limited inference. In: Notes in Computer Science: 6th Conference on Automated Deduction ed. D. W. Loveland. Springer-Verlag.

[Guha and Lenat, 1990] R.V. Guha and D.B. Lenat. Cyc: A Mid-Term report. AI Magazine, Volume 11, Number 3, 1990.

[Henderson, 1992] J. Henderson. A Connectionist Parser for Structure Unification Grammar. In Proceedings of ACL-92.

[Holldobler, 1990] S. Holldobler. CHCL: A Connectionist Inference System for Horn Logic based on the Connection Method and Using Limited Resources. TR-90-04S, International Computer Science Institute, Berkeley, CA.

[Kautz and Selman, 1991] H.A. Kautz and B. Selman. Hard problems for Simple Default Logics. Artificial Intelligence, 47(1-3), 243-279.

[Levesque, 1988] H.J. Levesque. Logic and the complexity of reasoning. Journal of Philosophical Logic, 17, pp 335-389.

[Mani and Shastri, 1992] D.R. Mani and L. Shastri. A connectionist solution to the multiple instantiation problem using temporal synchrony. In Proceedings of the Fourteenth Conference of the Cognitive Science Society. Lawrence Erlbaum.

[Mani and Shastri, 1991] D.R. Mani and L. Shastri. Combining a Connectionist Type Hierarchy with a Connectionist Rule-Based Reasoner. In Proceedings of the Thirteenth Conference of the Cognitive Science Society. Lawrence Erlbaum.

[McAllester, 1990] D.A. McAllester. Automatic recognition of tractability in inference relations. Memo 1215, MIT Artificial Intelligence Laboratory, February 1990.

[Miller, 1956] G.A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review, 63(2), pp. 81-97.

[Patel-Schneider, 1985] P. Patel-Schneider. A decidable first-order logic for knowledge representation. In Proceedings of 1JCAI-85. Morgan Kaufman.

[Rajasekaran, 1992] S. Rajasekaran. Personal communication.

[Selman and Levesque, 1989] B. Selman and H.J. Levesque. The tractability of path-based inheritance. In Proceedings of IJCAI-89. pp. 1140-1145. Morgan Kaufmann.

[Shastri, 1992] L. Shastri. Neurally motivated constraints on a working memory capacity of a production system for rapid parallel processing. To appear in the Proceedings of the Fourteenth Conference of the Cognitive Science Society. Lawrence Erlbaum.

[Shastri, 1991] L. Shastri. Why Semantic networks? In Principles of Semantic Networks. Edited by John Sowa. Morgan Kaufman Los Altos.

[Shastri, 1988] L. Shastri. Semantic networks : An evidential formulation and its connectionist realization, Pitman London/ Morgan Kaufman Los Altos. 1988.

[Shastri and Ajjanagadde, 1990] L. Shastri and V.G. Ajjanagadde. From Simple Associations to Systematic Reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. Technical Report MS-CIS-90-05, Department of Computer and Information Science, Univ. of Pennsylvania. (Revised January 1992). To appear in Behavioral and Brain Sciences.

[Ullman and van Gelder, 1988] J.D. Ullman and A. van Gelder. Parallel Complexity of Logical Query Programs. *Algorithmica,* 3, 5-42.