ANALOG: A Knowledge Representation System for Natural Language Processing

Syed S. Ali
Department of Computer Science
Southwest Missouri State University
901 South National Avenue
Springfield, MO 65804
ssa231f@cnas.smsu.edu

Background

The general task of work in knowledge representation and reasoning (KRR) is, trivially, the representation of knowledge. The domain for which the representation task has been undertaken is, often, secondary to the description of the overall task as an exercise in knowledge representation. Thus, the task of representing and using mathematical, common sense, visual, language-based, and logical knowledge is "lumped" together under the rubric of KRR (for examples of the diversity of goals of KRR see [12]). I will argue that in natural language processing and understanding there is a is a strong argument for the language structure and use mediating any attempt at representation. The task of knowledge representation for natural language processing and understanding is a knowledge-intensive one. To understand a natural language sentence a NLP system must, minimally, be able to represent the content of the sentence in the language of representation. These representation languages have been largely unmotivated with respect to the natural language they may be representing. In [12] only six of the twenty-two KRR systems presented are driven by natural language processing concerns.

Logic

Logic is a popular choice of representation languages because of is prefabricated syntax, and model-thoretic semantics. However, the syntax and semantics of first-order predicate logic is most definitely not that of natural language. I will argue that if the domain of representation is natural language, then the (minimally, syntactic) form of the natural language should strongly influence the representation.

If we contrast the various goals of researchers in the domains for which these kinds of knowledge are to be used, we may come to the conclusion that the goals differ significantly and, indeed, may conflict. Consider the traditional use of representations based on first-order logic. Here, we have a system that has a powerful, well-understood, inferential machinery but weak expressive power. Collections of logical formulas do not seem to capture the intuitive use of concepts by people. This representation for knowledge is unstructured and disorganized. What's missing is that first-order predicate does not provide any special assistance in the problem of what Brachman called "knowledge structuring" [6]. That is, the specification of the internal structure of concepts in terms of roles and interrelations between them and the inheritance relationships between concepts Attempts to incorporate knowledge-structuring into the representation language are typified by the use of frames, or frame description languages (FDLs). These are structured slot-filler structures representing concepts that stand in various relationships (i.e., taxonomic) to each other. However, this shift into slot-filler structures where the expressive power is greater but the inferential machinery is underspecified, perhaps, goes too far. The cost of knowledge-structuring is a weaking of the inferential machinery available.

Logic is a traditional representational medium for attempts to understand natural language. As a formal language logic has the desirable property that expressions of a formal language can be paired with interpretations. Thus a "correct" mapping from natural language sentences into logic allows the original sentences to be interpreted. However, the mapping from natural language structures into the standard syntax of logic (and from logic to natural language) is unnatural in that the structure of the natural language (which can be significant) is usually lost. This is typified by the "de-structuring" of noun phrases that occurs when it is translated into logic. This is a consequence of the atomic nature of variables in logic and results in the separation of constraints on variables from the variables themselves. This most often occurs when these constraints are moved to the antecedents of rules that type the variables. This results in the loss of the simple predicate-argument structure of the original sentence being represented.

Goals

This work is based on the assumption that natural language determines the design of a knowledge representation and reasoning (KRR) system. In particular, I present a KRR system for the representation of knowledge associated with natural language dialog. I will argue that this may done with minimal loss of inferential power and will result in an enriched representation language capable of supporting complex natural language descriptions, support for some discourse phenomena, standard first-order inference, inheritance, and terminological subsumption. Some of the goals of a more natural logic and its computational implementation in a knowledge representation and reasoning system are discussed below.

It should possess as much of the machinery of traditional FOPL as possible. It is not an accident that logical form representations are popular; it is a consequence of their power and generality.

The mapping from natural language sentences into logical form sentences should be as direct as possible, and the representation should reflect the structure of the natural language (NL) sentence it purports to represent. This is particularly well illustrated by rule-type sentences, such as "small dogs bite harder than big dogs," where their representation takes the form of an implication whose antecedent constraints specify what types of dog bite harder than another type. This representation, as a logical rule, contrasts with the predicate-argument structure of the original sentence, as below:

$$\forall x, y((\text{small}(x) \land \text{dog}(x) \land \text{large}(y) \land \text{dog}(y)) \\ \Rightarrow \text{bites-harder}(x, y))$$

By comparison, the representation of *Fido bites harder than Rover* is more consistent with the original sentence,

$$\text{bites-harder(Fido, Rover)}$$

This is so, despite the intutive observation that the two sentences have nearly identical syntactic structure, and similar meaning.

The subunits of the representation should be conceptually complete in the sense that any component of the representation of a sentence should have a meaningful interpretation independent of the interpretation of the entire sentence representation. For example, the sentence "dogs bite" is typically translated as:

$$\forall x[\text{dog}(x) \Rightarrow \text{bites}(x)].$$

With this translation, we might ask what is the meaning of x? Presumably, some thing in the world, or a set denoting the non-empty universe. Note that the original sentence mentions only dogs. We suggest that a better translation might be:

$$\text{bites}(\forall\, x \text{ such that dog}(x))$$

where the variable, x, has its own internal structure that reflects its conceptualization. Note that I am suggesting something stronger than just restricted quantification (the above could also be represented as $(\forall x : \text{dog}(x))\,[\text{bites}(x)]$ using restricted quantifiers). Complex internalized constraints (that is, other than simple type) and internalized quantifier structures characterize this proposal for the representation of variables. Thus representation of the sentence: *Every big dog that is owned by a bad-tempered person bites* should mimic the representation of *dogs bite*.

A high degree of structure sharing should be possible, as multi-sentence connected discourse often uses reduced forms of previously used terms in subsequent reference to those terms. This corresponds to the use of pronouns and and some forms of ellipsis in discourse. An example of this phenomena is the representation of intersentential pronominal reference to scoped terms, e. g.,

> Every apartment had *a dishwasher*. In some of them *it* had just been installed.
> Every chess set comes with *a spare pawn*. *It* is taped to the top of the box.

(examples from [10]). The structures that are being shared in these sentences are the variables corresponding to the italicized noun phrases. Logical representations can only model this "sharing" by combining multiple sentences of natural language into one sentence of logic. This method unnatural for at least two reasons. Firstly, when several sentences must be so combined into one sentence the resulting logical sentence is overly complex as a conjunction of several potentially disparate sentences. Secondly, this approach is counter-intuitive in that a language user can re-articulate the original sentences that he/she represents which argues for some form of separate representations of the original sentences. The problem with logic in this task is that logic requires the complete specification of a variable, corresponding to a noun phrase, and its constraints in the scope of some quantifier. This difficulty is not restricted to noun phrases, indeed it is frequently the case that entire subclauses of sentences are referred to using reduced forms such as "too" e. g.,

> John *went to the party*. Mary did, *too*.

A language-motivated knowledge representation formalism should model this sort of reference, minimally by structure sharing.

Any computational theory must incorporate knowledge-structuring mechanisms. In particular, subsumption and inheritance of the sort supported in frame-based and semantic network based systems. A taxonomy provides "links" that relate more general concepts to more specific concepts. This allows information about more specific concepts to be associated with their most general concept, and information filters down to more specific concepts in the taxonomy via inheritance. More general concepts in such a taxonomy *subsume* more specific concepts, the subsumee inheriting information from its subsumers. For atomic concepts, subsumption relations between concepts is specified by the links of the taxonomy. A clear example of subsumption in natural language is the use of descriptions such as *person that has children* subsuming *person that has a son*. If one were told: *People that have children are happy,* then it follows that *People that have a son are happy.* The intuitive idea is that more general descriptions should subsume more specific descriptions of the same sort, which in turn inherit attributes from their more general subsumers.

ANALOG

I have presented some general arguments for considering the use of a more "natural" (with respect to language) logic for the repesentation of natural language sentences. I have also presented some characteristics of natural language that a knowledge representation and reasoning system should support. In the full paper I will clarify the motivations for this work with specific examples, present an alternative representation for simple unstructured variables which involves according variables potentially complex internal structure, and reify the logic of these variables. This involves providing a syntax and semantics of the logic of structured variables. The syntax of the logic is specified by a complete definition of a propositional semantic network representation formalism (an augmentation of [14]). The implemented system is called ANALOG (A NAtural LOGic) [4, 5, 2, 1, 3].

The representations that that result from using these structured variables have the following advantage:

- The representation of numerous types of quantifying expressions, using structured variables, is more "natural" than typical logics, because the mapping of natural language sentences is direct. Thus parsing and generation of natural language sentences (particularly those involving restrictive relative clauses) is simplified.
- With structured variables that contain their own binding structures (quantifiers), no open sentences are possible. This is consistent with natural language where few, if any, open sentences occur (minimally, noun phrases are, by definition, typed).
- Sentences involving non-linear quantifier scopings, such as branching quantifiers and the donkey sentences can be represented [11, 9] because quantifier scope specification is internal to the structured variable.
- Subsumption can be defined syntactically it terms of the structure of structured variables. For example, from a rule representing *Every boy loves a girl* all rules involving more restricted boys follow directly by variable subsumption (e. g., *Every boy that owns a red car loves a girl*). This sort of subsumption can be defined without the

traditional T-Box and A-Box distinction (as in, for example, [7, 8]) and has a great deal of utility.

- Since the underlying representation is semantic network-based the representation of sentences that share structures (e. g., *Every farmer that owns a donkey beats it. His wife does too.*) is both possible and reasonable.

An example of the capabilities of the ANALOG system for NLP is given in Figure 1. ANALOG includes a generalized augmented transition network (GATN) natural language parser and generation component linked up to the knowledge base (based on [13]). A GATN grammar specifies the translation/generation of sentences involving complex noun phrases into/from ANALOG structured variable representations.

```
: Every man is mortal
I understand that every man is mortal.
: Who is mortal
Every man is mortal.
: Is any rich man mortal
Yes, every rich man is mortal.
: John is a man
I understand that John is a man.
: Is John mortal
Yes, John is mortal.
: Who is mortal
John is mortal and every rich man is mortal and every man is
mortal.
: Are all rich young men that own some car mortal
Yes, every young rich man that owns some car is mortal.
: Any rich young man that owns any car is happy
I understand that every young rich man that owns any car is
happy.
: Is John happy
I don´t know.
: Young rich John owns a car
I understand that mortal rich young John owns some car.
: Who owns a car
Mortal rich young John owns some car.
: Is John happy
Yes, mortal rich young John is happy.
```

Figure 1: Example of NLP in ANALOG.

References

[1] Syed S. Ali. A Propositional Semantic Network with Structured Variables for Natural Language Processing. In *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence.*, November 17-19 1993.

[2] Syed S. Ali. A Structured Representation for Noun Phrases and Anaphora. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 197–202, Hillsdale, NJ, June 1993. Lawrence Erlbaum.

[3] Syed S. Ali. Node Subsumption in a Propositional Semantic Network with Structured Variables. In *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence.*, November 17-19 1993.

[4] Syed S. Ali. *A "Natural Logic" for Natural Language Processing and Knowledge Representation.* PhD thesis, State University of New York at Buffalo, Computer Science, January 1994.

[5] Syed S. Ali and Stuart C. Shapiro. Natural Language Processing Using a Propositional Semantic Network with Structured Variables. *Minds and Machines*, 3(4):421–451, November 1993. Special Issue on Knowledge Representation for Natural Language Processing.

[6] Ronald J. Brachman. On the Epistemological Status of Semantic Networks. In N. V. Findler, editor, *Associative Networks: Representation and Use of Knowledge in Computers.* Academic Press, New York, 1979.

[7] Ronald J. Brachman, Richard E. Fikes, and Hector J. Levesque. KRYPTON: a Functional Approach to Knowledge Representation. *IEEE Computer*, 16(10):67–73, 1983.

[8] Ronald J. Brachman, Victoria Pigman Gilbert, and Hector J. Levesque. An Essential Hybrid Reasoning System: Knowledge and Symbol Level Accounts of KRYPTON. *Proceedings IJCAI-85*, 1:532–539, 1985.

[9] Peter Thomas Geach. *Reference and Generality.* Cornell University Press, Ithaca, New York, 1962.

[10] Irene Heim. Discourse Representation Theory, 1990. Tutorial material from ACL-90.

[11] W. V. Quine. *Philosophy of Logic.* Prentice-Hall, Englewood Cliffs, NJ, 1970.

[12] Charles Rich, editor. *Special Issue on Implemented Knowledge Representation and Reasoning Systems*, volume 2. ACM Press, June 1991. SIGART Bulletin.

[13] S. C. Shapiro. Generalized augmented transition network grammars for generation from semantic networks. *The American Journal of Computational Linguistics*, 8(1):12–25, 1982.

[14] S. C. Shapiro. Cables, Paths, and "Subconscious" Reasoning in Propositional Semantic Networks. In John F. Sowa, editor, *Principles of Semantic Networks*, pages 137–156. Morgan Kaufmann, 1991.