A Propositional Semantic Network with Structured Variables for Natural Language Processing

SYED S. ALI
Department of Computer Science
Southwest Missouri State University
901 South National Avenue
Springfield, MO 65804-0094, USA

ABSTRACT
This paper suggests some goals for knowledge representation for natural language processing: natural form, conceptual completeness, structure sharing. To address these goals, an augmentation to the representation of variables so that variables are not atomic terms is suggested. This leads to an extended, more “natural” representation language whose use and representations are consistent with the use of variables in natural language. It is shown how this representation simplifies the representation and use of complex noun phrases and resolves some representational difficulties with sentences involving nonlinear quantifier scoping, in particular, donkey sentences.

1 Introduction
The intent of the work in this paper is to specify a knowledge representation and reasoning formalism for natural language processing. To this end, I believe the following natural-language-specific goals must be addressed. First, the mapping from natural language sentences into the representation language should be as direct as possible, and the representation should reflect the structure of the natural language (NL) sentence it purports to represent. I call this the “natural form” constraint. This is particularly well illustrated by rule-type sentences, such as small dogs bite harder than big dogs, where the representation takes the form of an implication whose antecedent constraints specify what kind 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 structure of the original sentence,

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

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

Second, 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, for the representation of the sentence as in (1) above, we might ask what is the meaning of \( x \) or \( y \)? Presumably, some thing in the world, or a set denoting the non-empty universe. Note that the original sentence mentions only dogs. I suggest that a better
translation might be:

\[ \text{bites-harder}(\forall x \text{ such that small dog}(x), \forall y \text{ such that large dog}(y)) \]

where the variables, \( x \) and \( y \), have their own internal structure that reflects their conceptualization. Note that I am suggesting something stronger than just restricted quantification (simple type constraints can certainly be felicitously represented using restricted quantifiers). Complex internalized constraints (that is, other than simple type) and internalized quantifier structures characterize this approach to the representation of variables. Thus the representation of the sentence: *Every small dog that is owned by a bad-tempered person bites harder than a large dog* should reflect the structure of the representation of (2).

Third, 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 [7]). 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 is unnatural for at least two reasons. First, when several sentences must be combined into one sentence the resulting logical sentence, as a conjunction of several potentially disparate sentences, is overly complex. Second, this approach is counter-intuitive in that a language user can re-articulate the original sentences that he/she represents. This 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.

The donkey sentences [6] illustrate the utility of the goals of conceptual completeness and structure sharing. These are sentences that pronominally refer to quantified variables in closed subclauses, for example *Every farmer who owns a donkey beats it* where the noun phrase *a donkey* is a variable inside the scope of a universally quantified variable (*every farmer*) and is referred to pronominally outside the scope of the existentially quantified donkey. Figure 1 has some attempts to represent the above sentence in FOPL.

Representation (a) says that *every* farmer owns a donkey that he beats, which is clearly more than the original sentence intends. Representation (b) is a better attempt, since it captures the notion that we are considering only farmers who own donkeys; however, it contains a free variable. Representation (c) fails to capture the sense of the original sentence in that it quantifies over *all* farmers and donkeys, rather than just farmers that own donkeys. To see this, consider the case of the farmers that own two donkeys and beat only one of them. Clearly, the donkey sentence can apply to this case, but interpretation (c) does not. This is the interpretation I am interested in representing.

Summarizing, I have presented some general arguments for considering the use of a more “natural” (with respect to language) formalism for the representation of natural
language sentences. I have also presented some characteristics of natural language that a knowledge representation and reasoning system should support.

2 The Representation

I am attempting to represent variables as a “bundle” of constraints and a binding structure (quantifier). I term these bundles “structured variables” because variables, in this scheme, are non-atomic terms. The syntax and semantics of the associated logic is specified by an augmentation of a propositional semantic network representation formalism [9]. ANALOG (A NATural Logic) is described in [1, 3, 2, 4]. An example of an universal structured variable (the node labelled V1) is given in Figure 2. In Figure 3 node V2 is an existential structured variable node. The semantics of structured variables is largely unspecified, here, due to space limitations. It is an augmented (by the addition of arbitrary individuals) semantic theory based on [5, 10].

2.1 Natural Form

I suggest that the representation of numerous types of quantifying expressions, using structured variables, is more “natural” than typical logics. This is because the mapping of natural language sentences into representations is direct. An example is given in Figure 2.\(^1\) Note that the shaded node, labelled M1!, corresponds to the asserted proposition that all men are mortal. V1 is the structured variable corresponding to all men. The member-class case frame is the representation for the proposition that an object is a member of a class. This representation is more natural in that the top-level proposition is one of class membership, rather than a rule-like if-then proposition.

2.2 Conceptual Completeness and Structure Sharing

In typical logics, terms in one formula are not referenced in other formulas. In general, re-using a term involves re-writing the term in the new formula. Ideally, we would like to re-use exactly the same terms in different formulas, and we would like these terms to be closed formulas, i.e., complete, so that they may be meaningfully re-used and shared by multiple formulas. Variables with this property are called conceptually complete. This is also an issue in the representation of multisentential dialog, where intersentential reference to sententially scoped objects frequently occurs. For example, All cars come with a spare tire. It is not a full-sized tire, can only be represented by re-writing the terms

\[
\begin{align*}
(a) & \quad \forall x \ (\text{farmer}(x) \Rightarrow \exists y \ (\text{donkey}(y) \& \text{owns}(x, y) \& \text{beats}(x, y))) \\
(b) & \quad \forall x \ (\text{farmer}(x) \Rightarrow ((\exists y \ \text{donkey}(y) \& \text{owns}(x, y)) \Rightarrow \text{beats}(x, y))) \\
(c) & \quad \forall x \forall y ((\text{farmer}(x) \& \text{donkey}(y) \& \text{owns}(x, y)) \Rightarrow \text{beats}(x, y))
\end{align*}
\]

Figure 1: Three FOPL Representations of the Donkey Sentence

\(^1\) a node may be labeled by a “name” (e.g., BILL, M1, V1) as a useful (but extra-theoretic) way to refer to the node. This naming of a rule or proposition node is of the form Mn, where n is some integer. A “!” is appended to the name to show that the proposition represented by the node is believed to be true. However, the “!” does not affect the identity of the node, nor the proposition it represents. Similarly, variable nodes naming is Vn and base nodes naming is Bn where n is some integer.
corresponding to the shared variables in all representations of sentences that use them or combining several sentences into one logic representation. In these cases, there is a clear advantage to a conceptually complete (closed) variable representation (since there are no scope traps) as well as the structure sharing associated with a semantic network representation. With structured variables that contain their own binding structures (quantifiers), no open sentences are possible. Further, with structure sharing, distinct formulas may share constituents which would have been open formulas in a FOPL representation.

### 2.3 Donkey Sentences

Since there are no open subformulas, it is possible for formulas to use constituent variables at any level, including at a level which would result in an unscoped variable in a FOPL representation. Figure 3 shows the representation for *Every farmer who owns a donkey beats it*. Note that the shaded node, labelled $M_1!$, denotes the proposition *Every farmer who owns a donkey beats it* and $V_1$ and $V_2$ are *every farmer that beats a donkey he owns* and *a beaten donkey that is owned by any farmer*, respectively. This representation is also more natural because the top-level proposition is one of an act of beating, thus it is similar to any other sentence about beating such as *Fred beats a donkey*. This is the representation that is built and generated from when the donkey sentence is processed.
2.3.1 The Donkey Sentence Example

ANALOG includes a generalized augmented transition network (GATN) natural language parser and generation component linked up to the knowledge base (based on [8]). A GATN grammar specifies the translation/generation of sentences involving complex noun phrases into/from ANALOG structured variable representations.

Since this representational formalism is grounded in a nonlinear notation, the representation of tree-like quantifier scopings is straightforward. ANALOG's semantic network representation allows structure-sharing (and, indeed, may require it) to a high degree. One the initial goals was the representation of coreference and structure sharing in multisentential dialog. Because the parser/generator maintains a very simple discourse model, we do not illustrate complex examples (involving such multi-sentence connected discourse) of this structure sharing in natural language (although such representations may readily be built). A simpler example is the donkey sentence, where a scoped constituent of a noun phrase (the donkey in *Every farmer who owns a donkey beats it*) in used in the main clause of the sentence. Figure 4 illustrates a (slightly edited) dialog involving the donkey sentence.

![Figure 4: Example of structure sharing in donkey sentence.](image)

In Figure 4 the system reiterates its understanding of the donkey sentence with sentence (1). Note that the *some donkey* referred to in the subclause and main clause are the identical existential structured variable, the system cannot generate pronouns. The
system’s response with sentence (2) indicated that the representation of the donkey sentence is not that of Figure 1(c), since it would have answered yes had that been the case. The system’s response with sentence (3) indicates that the representation of the donkey sentence is not that of Figure 1(a), since the system just reiterates the rule rather than stating that Fred beats some donkey. Note that the system is initially unable to determine whether Fred beats Doc or Dumbo. This is because the initial rule (every farmer that owns some donkey beats it) is satisfied in a model where only one of the donkeys is being beaten. After the system is told that all such donkeys are beaten (sentence (4)), it does determine that Fred beats Doc. The answers to many questions are often rules themselves (e.g., who beats a donkey has as one answer Every farmer that owns some donkey). This is possible because the representations for rules have the same structure as ground propositions (a consequence the natural form and completeness constraints) and is illustrated in the remainder of the dialog.

3 Summary

I have presented some goals for a KRR system that would aid the process of NLP. I have described part of a KRR system that addresses these goals. ANALOG is a propositional semantic-network-based knowledge representation and reasoning system that supports many aspects of NLP, in particular, the representation and generation of complex noun phrases and sentences (natural form), the representation of various types of quantified variable scoping (conceptual completeness), and a high degree of structure sharing.

References