Semantic Analysis based on Ontologies with Semantic Web Standards

ARAKAWA Naoya
JustSystems Corporation
naoya.arakawa@justsystems.com

Abstract

This paper introduces an effort in the semantic analysis of natural language text based on ontologies and discusses the merits of using ontologies in semantic analysis. The semantic representation we use is based on discourse representation theory (DRT). As our ontology is written in OWL-DL and the semantic representation in RDF, the merits of using the Semantic Web standards will also be discussed.

1. Introduction

This paper illustrates an effort in semantic analysis of natural language text based on ontologies[1]. The distinctive feature of our semantic analysis is that its semantic representation is modeled with RDF graphs semantically disciplined with ontologies written in OWL. The reason for the use of RDF and OWL is that the Semantic Web technology has been emerging as the tool for representing and processing semantics and ontologies [1]. As it is crucial for semantic representation that its terms (elements) are defined in a definite way, ontologies are an indispensable part of semantic representation. Yet ontologies alone do not do the job; we need certain inference mechanism to derive meaningful results from semantic representation and ontologies. Here, the Semantic Web technology helps us with its representation standards (RDF as the general representation scheme and OWL as the language for representing ontologies) and its tools for inference.

The following sections describe how natural language semantics is represented in graphs, how semantic representation is related to ontologies or what kinds of ontologies are required for representing natural language semantics, how semantic analysis proceeds with syntactic analysis, lexicons and ontologies, and how semantic analysis is utilized for practical purposes.

2. Semantic Representation in Graphs

Our semantic representation draws most upon DRT (Discourse Representation Theory) [2] and SDRT [3] ("S" stands for "Segmented"). (S)DRT uses semantic representation called DRS (Discourse Representation Structure), which is a nested structure in which logical terms and predicates are placed. SDRT is a theory in dynamic semantics and purports to explain various discourse phenomena as well as semantics of simple sentences (and the versatility is the reason for our adoption).

Another feature of our semantic representation is that it has a graph representation. The reason for this is that we want our system to work with the Semantic Web, whose data representation model is in the graph form (i.e., RDF). While a DRS is not normally considered as a graph, there are theories that use semantic representations in graph forms. One example is *Conceptual Graphs* proposed by Sowa [4]. Conceptual Graphs represent a set of predicate logic formulas as a graph. Another example is CDL (Conceptual Description Language) proposed by Yokoi et al [5]. To be more specific, both Conceptual Graphs and CDL are hyper-graphs, i.e., nested graphs having sub-graphs as their nodes. Such nested structure or hyper-graphs are required, for example, for representing nested sentences.

---

[1] It is, of course, not in the sense of philosophy but that of artificial intelligence.

Now, as a DRS is nested structure containing predicate logic formulas, it can be converted into graphs in a fairly straightforward way. To represent predicate logic formulas in graphs, we can draw upon Conceptual Graphs. The nested structure of a DRS can be represented as a hyper-graph.

Here, we turn to the issue of representing hyper-graphs in RDF. As RDF is nothing but a graph representation, non-hyper-graphs can be represented in RDF with ease. There are at least two ways to represent hyper-graphs in RDF. In one way, a hyper-graph is represented in a single graph that includes sub-graphs and nodes representing the sub-graphs. A sub-graph is represented by a node representing the sub-graph itself, nodes within it and links between the sub-graph node and the nodes within. In another way, a hyper-graph is represented with a number of graphs and graph nesting is represented with a statement whose object is a graph ID (or URL). While the former is prima facie simpler (it only has one graph for representing a hyper-graph), the latter seems to be more straightforward way in terms of representing hyper-graphs. The former requires additional inference mechanism to suppress unwanted inference across sub-graphs (you shouldn't infer a fact from an expression in an 'opaque' context such as an embedded belief content). The latter also requires mechanisms to handle inference across sub-graphs. We have chosen the latter way with the hope that we can build inference mechanism for it based on pre-existing inference mechanism rather than making inference mechanism from scratch for the former way. There is also a public mechanism for supporting the latter way: SPARQL [6], the official querying language for RDF, supports 'named graphs' so that statements in a hyper-graphs can be retrieved.

3. Ontologies and Semantic Representation

An ontology in a broad sense is a set of rules for interpreting a system of symbols, or more practically, a set of inference rules with which machines draw inference over symbols. As semantic representation is a system of symbols, machines would need an ontology for drawing inferences. In contemporary AI, an ontology is the description of classes and relations in the domain of discourse. For example, an ontology may contain the description that an instance of the pencil class is an instance of the physical object class or color is an attribute taking a physical object as its subject. Inference mechanism uses descriptions in an ontology to draw or constrain inferences.

While inference based on ontologies can be used to draw implicit information from a given set of semantic representation in a knowledge base, it can also be used in semantic analysis or the process to create semantic representation from text. This is because semantic analysis uses rules and rules require generalization. For an ontology can contain a hierarchy of concepts, it provides with abstract concepts so that rules can be tersely written.

We have been developing ontologies for representing natural language semantics (currently not public) [6]. Its upper-most part is based on SUMO produced by IEEE [7][8], but we have intensively modified it. We are also compiling a large ontology for Japanese vocabulary of the size of 30,000, which is based on the EDR [9] concept dictionary. The ontologies are written in OWL-DL[2]. While the disciplined development of the large ontology requires the upper ontology, the latter has been heavily revised with the feedback from the development of the former, so that the entire development process makes a virtuous circle.

4. Lexicons

Semantic analysis requires a lexicon, i.e., a dictionary that links terms in text and their senses. A term sense is described as a concept in an ontology or as a semantic representation whose

[2] DL stands for description logic.

expressions are defined in terms of ontologies. This means that a lexicon requires an ontology.

In the area of lexicon study, the generativity has been an issue. A word may have a core meaning and derived meanings. Or, a word may be a morphologically derived from another with a derived sense from the sense of the original. For semantic analysis, we would need a lexicon that relates derived words and senses to their original so that the machine can tell the relation between the senses of terms used in text.

We have been developing a Japanese lexicon. Each term has multiple senses and each sense is currently linked to a concept in an ontology. The lexicon is generative in the following way to curtail unnecessary sense entries:

- Part-of-speech alteration is semantically invariant.
  The sense of the nominal form of a verb or adjective or the adverbial form of an adjective is the same as that of the original.
- The role holder concept such as 'teacher' is represented with its role concept so that the former can be dispensed with.
- Relational terms such as 'father' are represented with RDF properties[3] so that the corresponding classes such as the father class can be dispensed with.
- The meaning of the causative alteration of a verb is represented with a *CausalProcess* that causes the *StateOfAffairs* represented by the original verb.

5. Particulars of Semantic Representation and Semantic Analysis

In the following, I shall describe how semantic analysis proceeds with ontologies and lexicons. Concepts in our upper ontology (written in *italic*) are used for explanation. The process has been implemented on our prototype semantic analyzer, which converts dependency trees into RDF hyper-graphs. The implementation also gives feedback information for building the ontology. While the prototype is made for Japanese text, I shall use English examples for the sake of explanation.

5.1. Atomic Situations and Complex Situations

An atomic situation is a situation that a simple sentence represents. An atomic situation is represented with an RDF graph. States of affairs represented with sub-sentences in a coordinate sentence may be represented in an atomic situation if the sentence describes the same situation. A complex situation is a situation that a nested sentence represents. A complex situation is represented with a hyper-graph. (See figures in Appendix.) The translation of nested sentences to the representation of complex situations is mechanical conversion of nesting. A situation is represented as an instance of *IdeationalSituation* in our upper ontology. The ontology defines various logical and discourse relations between situations such as *precondition*, *consequent*, *entails*, and *example* (some of them are borrowed from SDRT). For each situation, veridicality and probability is calculated as its attribute. In a probable or opaque context, veridicality disappears.

5.2. Verbs and States of Affairs[4]

A state of affairs is what an argument structure of a predicate (verb) represents. In our semantic analysis, a verb is translated to an instance of the *StateOfAffairs* class and its case elements (arguments) to individuals that have semantic role relations with the *StateOfAffairs* instance. More specifically, a verb is associated with a subclass of *StateOfAffairs* via lexicon. The classification under StateOfAffairs is based on Beth Levin's [10] through SUMO, but revised as drawing upon LCS (the theory of Lexical Conceptual Structures) [11], adding classes such as

[3] An RDF property is a binary relation

[4] The author thanks Carol Tenny for useful discussions and advices on verb semantics.

CausalProcess. Our ontology defines semantic role properties such as *actor, origin, theme* and *method*, drawing upon SUMO, LCS, GDA [12], EDR and UNL [13], among others.

Tense is represented with properties representing temporal relations such as *before* and *after*. For example, past tense is represented as a situation whose occurring time is *before* that of the utterance situation.

Polarity is represented with the RDF property *truth* having the value (*TruthAttribute* instance) of *True* or *False*. In our implementation, *True* is the default while negation maps to *False*.

As for aspects, our ontology defines the following aspectual Attributes drawing upon EDR and Comrie [14]: Progressive, Completive, Habitual, Experiential, Persistent, Contingent and Resultative, and also Aktionsart classes as subclasses of *StateOfAffairs*: *State*, *Event* (Non-State), *TelicProcess* (change of state with a definite ending), *Accomplishment* (*TelicProcess* with a preceding durative phase) and *SemelfactiveEvent* (momentary Event). The aspect of verbs has complicated interaction with the Aktionsart. For example, while Japanese verbs representing telic processes in the '-teiru' form usually have the resultative aspect, those representing atelic processes with the same ending usually have imperfective aspects.

As for modality, our ontology defines the following *AlethicAttribute*s such as *Necessity, Impossibility* and *Possibility* and *DeonticAttribute*s such as *Obligation* and *Permission*, as well as modal properties such as *requires*.

5.3. Nouns

A noun is by default translated to an individual (an instance of owl:Thing) with additional information from the lexicon. Proper names (e.g., "Cleopatra") in most cases are translated to the RDF resources representing the individuals the nouns denote (e.g., Cleopatra) or simply to individuals having the name strings. Most common nouns (such as 'animal') are also translated to individuals but typed with classes (such as *Animal*) via lexicon. As for nouns representing roles such as "teacher," special care is taken, for our ontology unifies role concepts and the concept of role holders to reduce redundancy. When a noun is associated with a role class, the generated individual representation is not directly typed with the class but linked to an instance of it with the property *role*. (Fig. 1) Similarly, a noun may be associated with an RDF property such as *father*. In such a case, the generated individual representation is linked to the attribute as either its subject or object (the choice is made by information in the lexicon or other resources). (Fig. 2)

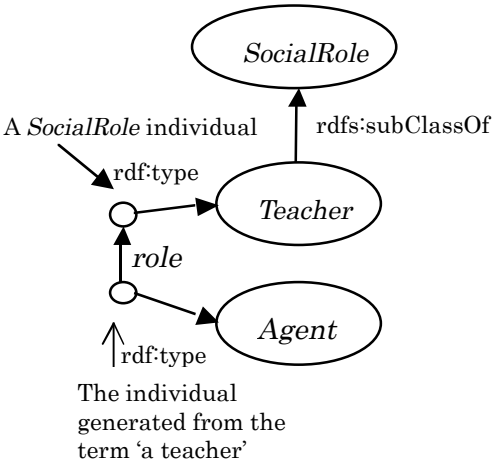


Fig 1.

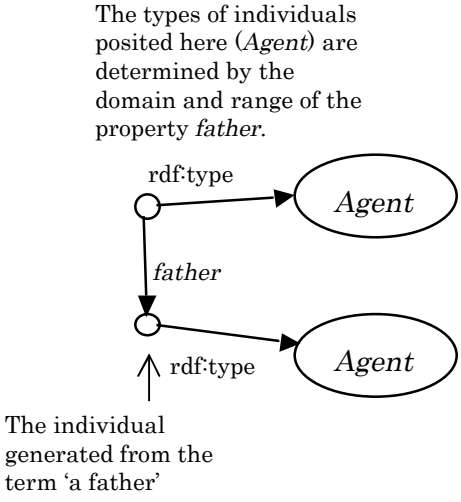


Fig 2.

5.4. Plurality

A plural noun is translated to an instance of *Collection*. For example, "people" is translated to an instance of *GroupOfAgents*, a subclass of *Collection* whose members are defined to be

instances of *EmbodiedPerson*. A coordinate structures of nouns with the conjunction "and" is also by default translated to an instance of *Collection.*

5.5. Adjectives and Adverbs

Adjectives and adverbs are normally translated to an instance of the abstract class *Attribute*. Both predicative and attributive use of an adjective is translated to an individual linked to the instance of *Attribute* with the RDF property *attribute*. Gradable adjectives and adverbs are translated to instances of *GradableAttribute*, which is a subclass of *Attribute* and comparative expressions are expressed with comparative relations such as *lessThan* placed between instances of *GradableAttribute*.

5.6. Modification

As mentioned earlier, the modification by adjectives and adverbs is represented with the property *attribute*. Generic modification with the preposition "of" (or the postposition "*の*" in Japanese) is represented with the property *related*. Modification with more specific prepositions is translated to appropriate sub-properties of *related*.

5.7. Quantification

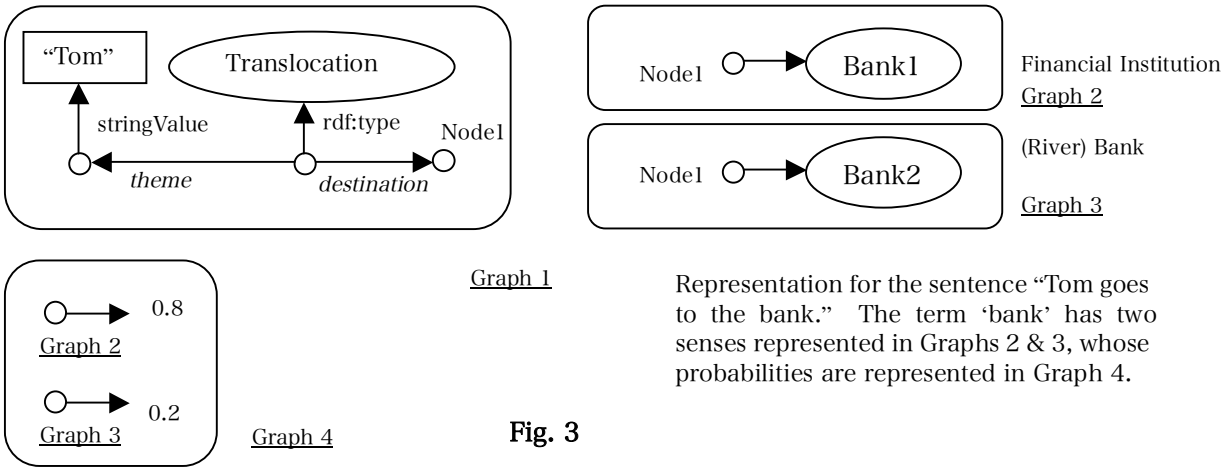
Logical quantification requires quantifiers and nested representation of scope. Our semantic representation regards existential quantification as default and for universal quantification, the attribute *All* is given to the individual to be quantified. Conceptual Graphs use similar treatment for quantification. Our ontology defines other quantifiers such as *This*, *That*, *Most* and *Generic*. Nested scopes can be represented with nested situations, which are in turn represented with a hyper-graph. To note, scopes do not always have to be specified (underspecification).

5.8. Utterances

A sentence is translated to an instance of *Utterance*. *Utterance* is a subclass of *Expressing* and *LinguisticCommunication*, which are subclasses of *IntentionalProcess*. As an *Utterance* is a *StateOfAffairs*, it takes semantic roles such as *actor*, and the situation where an *Utterance* occurs can have properties such as *time* (see Fig 4.3 in Appendix). *Utterance* has subclasses such as *Ordering, Questioning, Requesting* and *Declaring* as its speech act types. For setting these types, we have drawn upon FIPA performatives [15], GDA communicative functions and attributes in the EDR corpus.

6. Ambiguity

In our semantic representation, each sense of ambiguity is explicitly represented with an RDF graph. For example, the senses of the word "bank" registered in the lexicon are placed in separate RDF graphs from the main graph where the individual generated from the noun is placed (Fig. 3). Each graph can have its probability given by the lexicon or calculation based on collocation.



Representation for the sentence "Tom goes to the bank." The term 'bank' has two senses represented in Graphs 2 & 3, whose probabilities are represented in Graph 4.

Fig. 3

Recent theories of semantic representation such as MRS [16] and SDRT (glue logic) have the mechanism for underspecification, where ambiguity is left implicit. With our graphic semantic representation, underspecification can be realized in a straightforward way. For example, dependency relations or discourse relations can be underspecified just by not adding a graph edge (RDF property) representing these relations.

7. Related Works

There are various attempts to translate (controlled) human languages into semantic representation. Some of them are based on ontologies. The table below is a non-exhaustive list of such systems.

System	Source Language	Sem. Reps.	Base Ontology Format
CLCE [17]	Controlled English	FOL	-
ACE [18]	Controlled English	DRS	-
UNL [19]	Various Languages	UNL	-
NKRL [20]	English	NKRL	Original
Cyc [21]	English	CycL	CycL
Ours	Japanese	DRS/RDF	OWL-DL-based

Table 1

In comparison, the characteristics of our system are summarized as follows.

- The source language is not controlled.
- The analysis is based on independent ontologies written in OWL-DL.
- The ontologies draw on linguistic theories and are moderately formal.

8. Further Development

This last section discusses things to be done for better semantic analysis. While we have built a basic mechanism for building DRT-like representation from Japanese text, the first issue to be tackled is the accuracy of analysis. Even if the dependency analyzer gives the correct result, the semantic analyzer may not produce correct representation. One source of errors is semantic role assignment, which arises from semantic irregularity in the relation among assigned semantic classes for verbs, semantic roles and case markers. Another source is idiomatic expressions. These irregularities would be handled by hand-coded rules with ontological terms.

To draw factual information from text, one thing the analyzer must do (besides textual inferences) is to resolve anaphors. In Japanese, pronouns are often omitted (ellipsis) so that the system should recover 'zero pronouns' before resolving the anaphor. Our linguistic ontology is useful for recovering pronouns and resolving anaphor, for it has the information on the obligatory semantic roles for *StateOfAffairs* subclasses and the categories (range) of semantic roles (the information is coded with owl:someValuesOf statements). The system can recover omitted semantic role elements for a verb by associating it with the obligatory semantic roles of associated verb classes. The ontology helps resolving anaphor by providing semantic categories for anaphors and their antecedents to be matched.

Finally, I give a few words on practical side of this research. Currently, we are developing a semantic search system combining the semantic analyzer and public domain tools for semantic repositories. Such a system may serve as a new kind of text retrieval tools and generate good input for text mining.

References

[1] A Semantic Web Primer, G. Antoniou & F. v. Harmelen, MIT Press, 2008.

[2] *Handbook of Logic and Language*, J. F. A. K. Van Benthem, et al. (eds.), MIT Press, 1997.

[3] N. Asher and A. Lascarides, *Logics of Conversation*, Cambridge University Press, 2005.

[4] Conceptual Graphs : http://conceptualgraphs.org/

[5] ISeC CDL Technical Reports：
http://www.instsec.org/tr/

[6] N. Arakawa, "The Trial Development of Upper and Large-Scale Ontologies for Natural Language Processing", SIG-SWO-A602-04, Japanese Society of Artificial Intelligence:
http://www.jaist.ac.jp/ks/labs/kbs-lab/sig-swo/papers/SIG-SWO-A602/SIG-SWO-A602-04.pdf, 2006 (in Japanese).

[7] IEEE Standard Upper Ontology WG SUMO home: http://suo.ieee.org/SUO/SUMO/

[8] A. Pease, SUMO home: http://www.ontologyportal.org/

[9] EDR home: http://www2.nict.go.jp/r/r312/EDR/index.html, http://www2.nict.go.jp/r/r312/EDR/index.html

[10] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigation*, Univ. of Chicago, 1993.

[11] R. Jackendoff, *Semantic Structures*. MIT Press, 1990.

[12] K. HASIDA, GDA tag set,: http://i-content.org/GDA/tagset.html

[13] UNL Specifications：
http://www.undl.org/unlsys/unl/unl2005/

[14] B. Comrie, *Aspect*, Cambridge University Press, 1976.

[15] FIPA Communicative Act Library Specification:
http://www.fipa.org/specs/fipa00037/SC00037J.pdf

[16] A. Copestake, et al., "Minimal Recursion Semantics - An Introduction":
http://www-csli.stanford.edu/~aac/papers/newmrs.pdf

[17] J. Sowa, "Common Logic Controlled English," 2004: http://www.jfsowa.com/clce/specs.htm

[18] N. E. Fuchs, et al. "Discourse Representation Structures for ACE 6.0" 2008:
http://attempto.ifi.uzh.ch/site/pubs/papers/drs_report_6.pdf

[19] H. Uchida, et al., *Universal Networking Language*, 2005: http://www.undl.org/

[20] G. P. Zarri, "NKRL, a knowledge representation language for narrative natural language processing," COLING 1996 Proceedings, pp. 1032-1035, 1996.

[21] Cyc-NL: http://www.cyc.com/cycdoc/ref/nl.html

Appendix

Semantic representation of the sentence “Tom believes a cat chased a hen and a duck” in RDF. (Actual analysis was done with the equivalent sentence in Japanese. Class names in the figures have been translated into English.)
Fig 4.1 is a graph showing Tom’s believing the IdeationalSituation HID03 as the content.
Fig 4.2 is a graph showing the content of the belief. The past tense is indicated with the property GUO:before and the time of utterance /EID1_T defined in Fig 4.3. The conjunction “a hen and a duck” is represented as a collection.

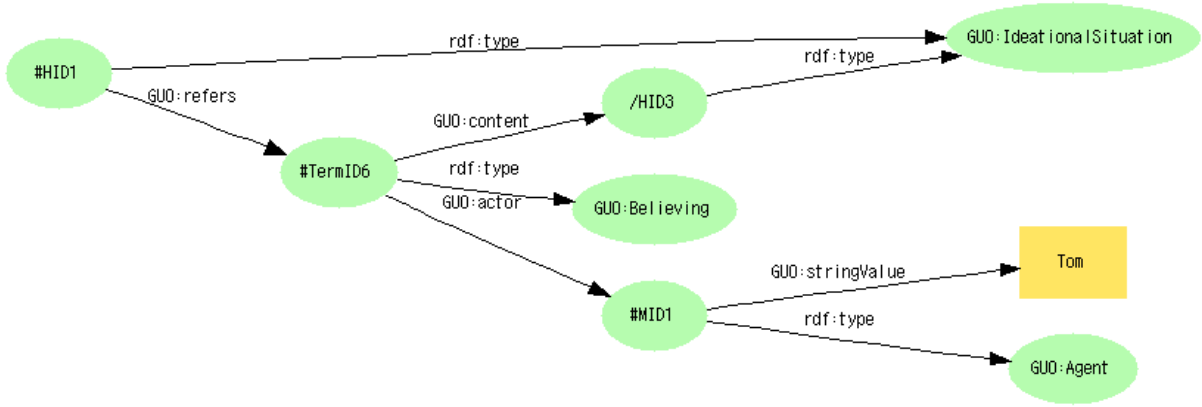


Fig 4.1 Main Graph

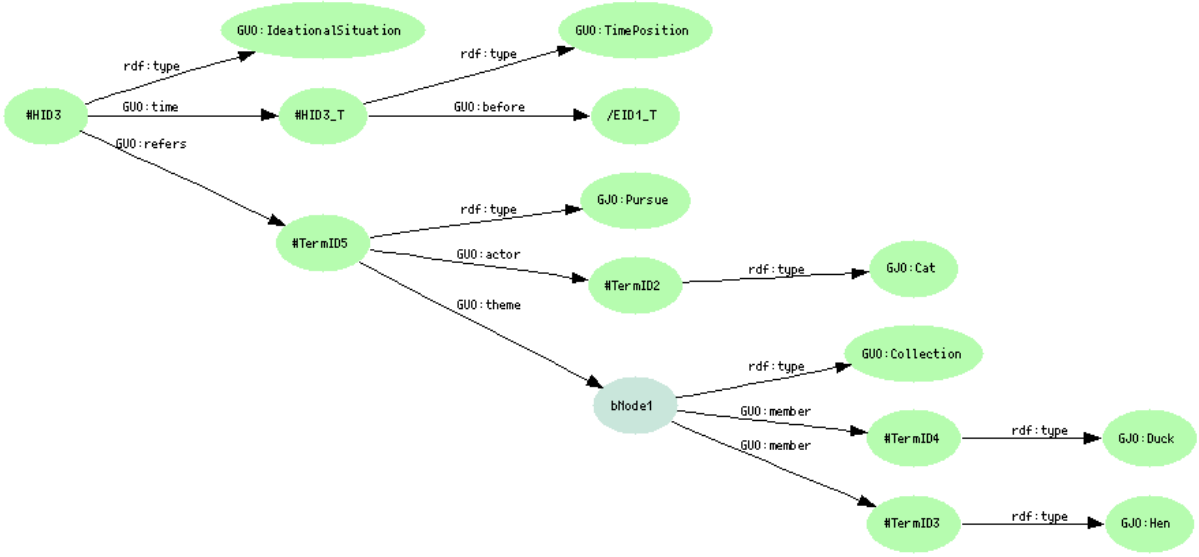


Fig 4.2 Belief Content

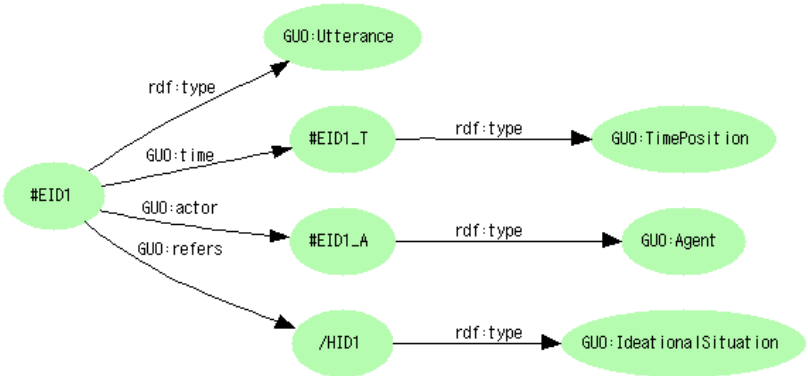


Fig 4.3 Utterance