Modelling Knowledge for a Natural Language Understanding System

Gudrun Klose, Thomas Pirlein

IBM Germany
Scientific Center
Institute for Knowledge Based Systems
P.O.Box 80 08 80
D-7000 Stuttgart 80
Email: KLOSE@DS0LILOG, PIRLEIN@DS0LILOG

Abstract

In the field of knowledge based systems for natural language processing, one of the most challenging aims is to use parts of an existing knowledge base for different domains and/or different tasks. We support the point that this problem can only be solved by using adequate metainformation about the content and structuring principles of the representational systems concerned. One of the prerequisites in this respect is the transparency of modelling decisions.

After a short introduction to our scenario, we will propose *general dimensions* for characterizing knowledge in knowledge based systems. These dimensions will be differentiated according to linguistic levels of investigation in order to deduce *structuring principles* for the modelling process. The resulting criteria will be evaluated in a *detailed example* taken from our prototypical implementation.

We hope to contribute some promising steps towards a methodology of knowledge engineering with natural language and common sense orientation.

1 Introduction

In the following, we want to sketch first results of knowledge engineering research which was undertaken for the LILOG project (Linguistic and logic methods). LILOG develops concepts for natural language systems for text understanding. Major results are available in a prototype system LEU/2[1] (LILOG Experimentier- Umgebung)[2].

In order to reduce the complexity of the system, it has to be decomposed into modules.

Our approach embodies modules oriented towards levels of linguistic investigation like morphology, syntax and semantics. In addition the modules differentiate between analysis and the generation processes. In the ideal case, all processes and modules will be supported by commonsense knowledge.

A crucial problem in this context is the construction of an adequate background knowledge base. The need for a methodology is obvious. First steps have been made in expert system research, where both domain and task are for the most part clearly specifiable. This does not hold for systems with natural language - and common sense orientation. In what follows, we will outline the knowledge engineering approach in LILOG along three dimensions.

Task:

Domain and texts were selected in order to cover a wide variety of linguistic phenomena to be handled by the linguistic parts of the system (i.e. parsing and generating components). In order to prove the appropriate understanding of the texts, the architecture was designed a.o. as a question/answer system. Hence, we get the additional task to generate language.

Domain:

For LEU/2, the domain was restricted to travel guide information about the city center of Düsseldorf. As a first step, a set of written data was obtained by travel guides, supplemented by travel agencies and a local inspection of Düsseldorf city center.

The set of different entities was to meet the following conditions: it should be large enough for a relevant size of the knowledge base, interconnected enough to allow for interesting inferences but at the same time small enough for being handled within a prototypical implementation.

We decided to work with a couple of short texts (frequently found in travel guides), which describe

[1] Leu/2 is being developped at IBM Germany in cooperation with some university partners, and is fully implemented in Prolog under AIX. The knowledge base for the domain under investigation consists of about 600 concept definitions, among these some 100 belonging to the upper structure. The number of attributes for each of these concepts averages around 20. At this time the number of axioms for our domain is approximately 300.

[2] "LILOG Experimental Environment"

particular sightseeing items, and a one page narrative text about a group of people on a prototypical sightseeing tour. In the next step, the chosen texts were classified according to linguistic criteria and analyzed for their propositional contents.

Granularity:

In order to obtain a first hint at the variety of text understanding tasks which LEU/2 was intended to deal with, native speakers were asked to formulate questions and to provide acceptable answers concerning the contents of the texts.

The selection of items and the way these native speakers talked about them, served as guideline to determine an appropriate granularity of the knowledge base.

The overall performance of the system is determined by the interaction of its *components*. Due to the modular approach, the relevant subtasks of the *knowledge base* had to be separated from those of the *lexical, syntactic, semantic analysis components* and the *generation module*. As a result of this preliminary investigation, three **dimensions of knowledge** turned out to be crucial to the modelling process.

2 Dimensions of Knowledge

We will discuss knowledge from two different perspectives. On the one hand we have those conditions which lead to qualitative requirements concerning the contents of the knowledge base. The other perspective concerns aspects induced by formal devices, i.e. the knowledge representation formalism used.

2.1 Qualitative Dimensions

If you consider knowledge representation as a special case of model theory, you will get a hint of how to proceed. As to the *breadth* of the model, the first dimension at issue, this means:

> The job of the representing world is to reflect some aspects of the represented world in some fashion.[Palmer, 1978]

As regarding *granularity*, the second dimension, a model reflects only a subset of the characteristics of the entities it represents. This, in turn, determines the depth of the model.

A third dimension is given by the *complexity of the task* the model is intended to cover.

All three dimensions are shown in picture 1.

Some of the consequences for the model in LILOG following from this view of knowledge representation are described below.

2.2 Formal Devices of Representation

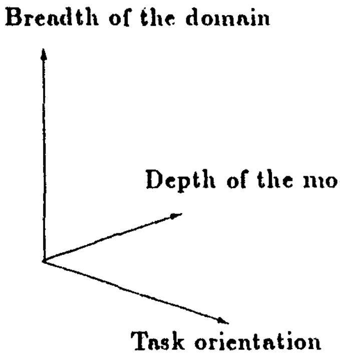


Figure 1: Qualitative dimensions of knowledge

In the field of logic based formalisms for coding background knowledge in natural language processing systems, there is some controversy on the design and use of formal constructs. Topics in this debate are the function of axioms compared to recent expert system technology, the function of structured concept hierarchies [Monarch and Nirenburg, 1987], the quality and number of additional attributes (roles in KL-ONE like systems) or syntactic validation criteria [Horacek, 1989]. Our approach aims at finding useful selectional criteria for different expressive means of the formalism $\mathrm{L}_{\mathrm{LILOG}}$ in order to bridge the actual gap between *problem driven* and *technology driven* [3] research.

We can make use of two kinds of formal constructs:

- A frame-description language similar to KL-ONE (cf. e.g. [Brachman and Schmolze, 1985]), which serves to represent the terminology of the domain by means of
  - sort expressions for classes of entities, organized hierarchically as sets and subsets (i.e. the logical subsumption relation), and
  - two place predicates and functions (i.e. features and roles), attached to specific sorts and constituting functional and relational connections between sorts, and
- axioms of first order predicate logic, expressing inferential dependencies between domain terms in form of the axiomatic semantics for those terms.

So the formalism used here[4] is comparable to e.g. KRYPTON (s.e.g. [Brachman et al., 1985]).

In the following, we will discuss the qualitative dimensions of knowledge in more detail. We will focus the qualitative criteria by differentiating them according to our scenario.

[3] See [Lehnert, 1988] for that distinction.

[4] For a detailed description of the formalism $\mathrm{L}_{\mathrm{LILOG}}$ see [Pletat and von Luck, 1989]

3 Criteria for Structuring the Ontology

3.1 Demands Resulting from the Task

As mentioned above, the task of our system is to simulate text understanding. This requires a transfer of insights from linguistic research into knowledge engineering. In the ideal case, structures of the model will be strongly influenced by natural language analyses.

Linguistic knowledge is relevant in various respects:

- **Word orientation**, for example, implies close interrelationships with research on lexical knowledge: affiliated generic terms, discriminating features, idiosyncratic aspects of use, etc. However, you may run into difficulties by relating syntactic categories (like word classes) with conceptual structures. So thematic roles cannot be directly transformed into ontological roles as a part of the background knowledge. In the sentence

  > The bus took the participants of the conference to the city center.[5]

  the 'bus' is an agent of an event from the syntactic point of view and at the same time conceptualized as instrument (and not agent) of an event in an ontological sense.

- **Sentence oriented** linguistic investigation implies the reconstruction of knowledge on the sentence level, as opposed to the meaning of single words or of textual structures. As an illustration might serve temporal information about the progress of actions or situations. Theoretical work in this field was initiated e.g. by Z. Vendler [Vendler, 1967] with his analysis of verbs and times. His differentiation of *states*, *activities*, *accomplishments* and *achievements* has been established as a well known classification of verbs. One important criterion for this distinction is the *goal-orientedness* of the concerned verbs: states and activities are by definition not goal-oriented, whereas accomplishments and achievements are goal-oriented in a temporally extended or punctual way, respectively.

  The aspect of goal-orientedness turned out to be central in our domain, e.g. as to directional verbs of movement. The sentence

  > The tourists took the bus to the Rhine and went for a boat trip.[6]

  allows to infer by default that the tourists reached their goal (the Rhine), because the location of the following event (the boat trip) is the same as the arrival point of the bus ride. By introducing goal-orientedness as a part of the definition of events, it will hence be possible to give an affirmative answer to the question

  > Were the tourists at the Rhine?[7]

- Moreover, a text necessarily involves **discourse oriented** information. Text understanding phenomena like anaphora resolution can only be accounted by accessing background knowledge concerning interconceptual relation.

  > The tourists went for a boat trip. They took the seats on the sundeck.[8]

In order to capture the meaning of these sentences, three steps have to be inferred: A boat trip is usually undertaken with a boat; a boat often has a sundeck; and a sundeck mostly offers seats.

3.2 Demands Resulting from the Domain

In the LEU/2 context, we have to deal with the comprehensive task of text understanding and a relatively narrow domain. Consequently, the general problem of conceptualization is limited by a restricted number of entities relevant to our field. Modelling these entities includes both the selection of concepts which appear in the domain, and the plausible combination and summing of recurrent concepts. The plausibility of modelling decisions in this sense can be judged from an *engineering* point of view in terms of optimizing search space (system performance) and from a philosophical point of view in terms of the principle of *economy of the ontology*.

The concepts RESTAURATION, CONSTRUCTION and RENOVATION may serve as an illustration taken from our domain. As they share similar aspects and inferences, we decided to introduce the supersort MODIFICATION (see section 4).

3.3 Granularity: Depth of Modelling and Inferencing

In the third qualitative dimension of knowledge we have to face the problem of delimitating the depth of the model in order to reduce complexity. As it is not possible to give

[5] The German version of the sentence is part of the text corpus of LEU/2: "Der Bus brachte die Teilnehmer der Konferenz in die Innenstadt".

[6] "Die Touristen nahmen den Bus bis zum Rhein und machten einen Bootsausflug."

[7] "Waren die Touristen am Rhein?"

[8] The German version of the sentence is part of the text corpus of LEU/2: "Die Touristen machten einen Bootsausflug. Sie nahmen die Plätze auf dem Sonnendeck ein".

an exhaustive system of categories[9], it seems legitimate to determine primitive concepts dependent on the chosen task and domain. In addition, selectional criteria for clusters of inferences have to be determined. (See example in section 4). As a possibility of measuring the depth of a model, Hayes ([Hayes, 1979]) proposed a ratio of axioms per concept.

Aside from measuring the expression of dimensions of knowledge by means of quantitative data, it is important to consider qualitative dependencies between the *depth* and *task* of the model on the one hand and between the *depth* and *domain* on the other.

Depth in relation to the task

Within the task of text understanding, some requirements of representation are e.g. goal orientation, culmination, causal connections, intention, etc. [Trabasso and Sperry, 1985]. In all these cases the chosen granularity has strong impact upon the resolution of interrelations in the texts.[10]

Depth in relation to the domain

This connection can be illustrated by the following example: A typical event of our domain is RESTAURATION. In our scenario, touristic aspects like the architect (agent), the time and the object concerned (e.g., the facade) will be of crucial importance. Given a different scenario like the protection of historical monuments, we would have to face an interest in considerably more details, requiring the choice of a deeper granularity.

4 Design of the Knowledge Base

In this section, we first want to give a brief survey of the ontology. After that, we will take up the sorts and regularities mentioned so far and present a structured exemplary model formalized in L_{LILOG} .

Sort expressions are used to represent the categories of our domain model. The *upper structure* of the resulting ontology portrays some generalized schemes of organization of relative domain-independence. When descending the model towards the *lower structure*, the categories are defined much closer to the word level and therefore domain-specific in the sense of *explicit* text knowledge.[11]

As already mentioned, we want to simulate understanding of basically two different types of texts, i.e. short texts describing single sightseeing items and narrative texts dealing with sequences of events. This lends us to the requirement of both an object-oriented and an event-oriented part of the conceptual hierarchy.

Consequently, one of our basic design decisions is due to J. Hobbs (cf. [Hobbs et al., 1987]) and results in a reification of predicates. So in our model all events, states etc. have concept status on their own.

This technique enables us to model the case frames for verbs in an analogical manner to the lexical entries of the analyzing component as well as to incorporate the structures for events etc. within the categories alike the definitions for objects.[12] It makes sense to think about objects as well as about events in terms of their spatial and temporal environment, although these knowledge specifications will obviously be quite different.

An example taken from the event cluster may serve as an illustration of several consequences of the criteria mentioned above. As to the **breadth** of the model, the relevance of the event part of the ontology appears intuitively plausible with respect to our domain, namely a scenario of cities, with modifying events. We have to deal with sights of the city like facades of important buildings, and the events of modification related to them show a considerable resemblance of important features of meaning – although the verbs are no real synonyms in the linguistic sense.

Figure 2 shows a screen dump with the relevant part of the concept hierarchy. The picture illustrates the effect of bundling that the introduction of adequate superconcepts has, and which allows for structured inferencing in terms of system efficiency. In this part of our concept hierarchy the boarderline between Upper Structure and Lower Structure is clearly identifiable. When descending the hierarchy, the sort KONSTRUKTIVSIT fans out into several domain-dependent subsorts.

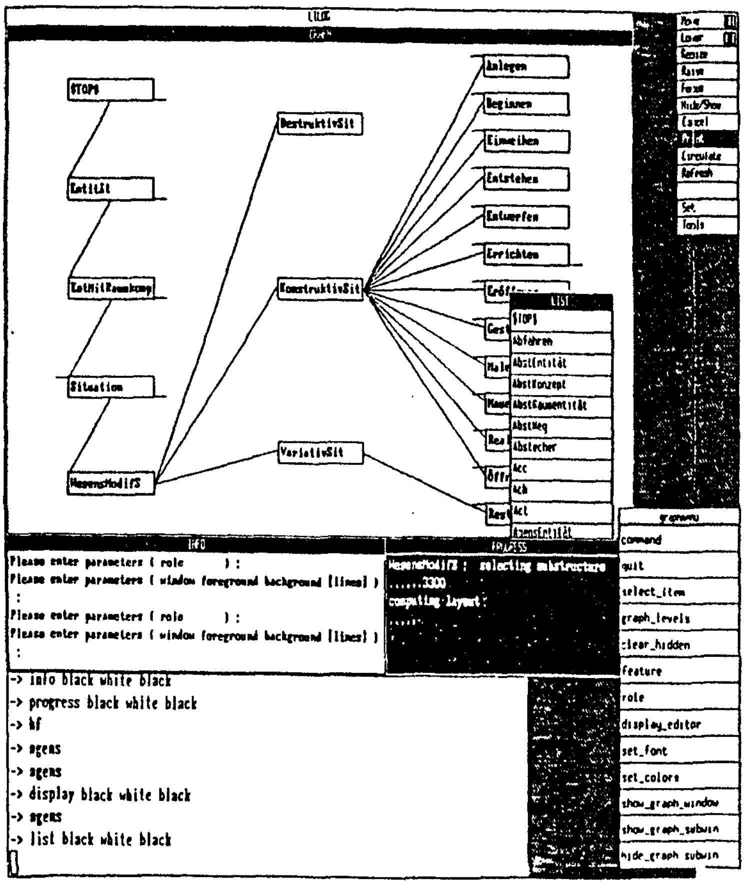
The figure is followed by the respective sort expressions written in the L_{LILOG} list structure(the sort KONSTRUKTIVSIT in the figure corresponds to CONSTRUCTION in the English list of sort expressions), expanded by roles and features which do not appear in the graphic representation. It should be noted here that a third kind of information is omitted even in the list notation. More general roles and features (like e.g. agent, time and so on) are inherited by superconcepts and not visible in neither presentation. (The short line in the upper left corner of some concept boxes indicate the existence of additional hidden superconcepts.)

[9] See for example [Tamas, 1986, p. 509]

[10] For a more detailed discussion, see [Pirlein, 1990].

[11] This differentiation between *upper* and *lower* structure of the model is introduced by [Mann et al., 1985].

[12] A similiar technique you can find e.g. in [Mann et al., 1985].



```
Please enter parameters ( role         ) :
Please enter parameters ( window foreground background [lines] )
 :
Please enter parameters ( role         ) :
Please enter parameters ( window foreground background [lines] )
 :
```

```
WesensModifS :  selecting substructure
......3300
computing layout :
.....
```

```
-> info black white black
-> progress black white black
-> hf
-> agens
-> agens
-> display black white black
-> agens
-> list black white black
```

Figure 2: Implementation of the 'modification'-event

The definition of the relevant event concepts in L_{LILOG} is followed by an axiom which transfers information about the time of a construction event to the beginning of lifetime of the concerned object. This kind of structured modelling allows to dispense with writing similar axioms for a number of resembling events.

In order to demonstrate task orientation, it would be necessary to consider a broader part of the ontology, because aspects like intention, causality or culmination have been modelled separately. In addition, one would have to take a closer look at the ensemble of connected components in the system. The limitation of the depth of the model can be seen from the fact that the event concepts discussed do not have more differentiated subconcepts and, of course, from the fact that not all possible roles and features have been integrated into the model. In a scenario "protection of historical monuments", for example, the instruments of renovation might be central and would induce a partly different granularity in the model.

Definition in L_{LILOG}:

```
sort modification          < situation.
sort essential modif       < and(modification,
                                  essential_obj :
                                    object).

sort variation             < essential_modif.
sort construction          < essential_modif.
sort destruction           < essential_modif.
sort material variation    < and(variation,
                                  essential obj :
                                    material object).

sort restauration          < material variation.
sort material construction< and(construction,
                                  essential obj :
                                    material object).

sort mental construction   < and(construction,
                                  essential obj :
                                    not(material object).

sort building              < material construction.
```

The twofold modelling of PHYSICAL and MENTAL CONSTRUCTION is e.g. necessary to distinguish ideas developped by an architect from the realization of the building.[13]

For constructive events one can define the following regularity (axiom):

```
axiom rule30  forall D1 :  construction,
                     02 :  object,
                     T3 :  timeintervall;

                        essential obj(D1,02)
                     and livetime(02,T3)
                     ->
                     meets(D1,T3).
```

The relation meets is one expression of our axiomatization of Allen's time interval logic [Allen, 1983] in L_{LILOG} . Rule30 exemplifies a transformation rule between the clusters of events and objects, respectively.

Our task setting implies certain ways of interaction between Knowledge Engineering and the generation component. If you want to obtain flexibility for the generation component with respect to the possible diversity of answers, information should be available in cases of object centered questions

("What do you know about object xy ...")

as well as in comparable event oriented requests

("What happened after ...").

[13] For reasons of clarity we renounced on showing all respective supersorts.

5 Conclusion

One of the most discussed topics in the field of text understanding is the separation between semantic knowledge on the one hand and common sense knowledge or world knowledge on the other. During the conception and implementation of the modules in our prototype, this discussion was reflected by a considerable flexibility in the division of functions between semantic analysis and inferential processes.

During the integration, descriptive parts of linguistic theories had to be completed with procedural or functional aspects. Typical misfits appeared each time it was clear *what* should be *expressed* within certain modules (like morphology or syntax), but it was unclear *how* to *proceed* from one module to the next. In the ideal case, this allowed for conclusions on incompatibilities between the levels of linguistic analysis corresponding to the respective modules.

One of these phenomena is the identification of adjectival passive constructions versus regular verb:

> The museum will be opened at 11 a.m.[14].
>
> The museum is open from 9 to 15[15].

According to Vendler's classification, *open* should be categorized as an event in the first sentence and, combined with *to be* in the second case, as a state. The integration of the modules showed that none of the system components was able to deliver this differentiation - in this case, the reason was the incompatibility between unsorted unification grammars and the necessity to overwrite default values.

In the field of Knowledge Engineering, the question how to make contents of one knowledge base available to a second one (normally with quite another kind of task setting) has been receiving growing attention. One of the most interesting parts of this problem consists in the interrelationship between common sense - and domain specific knowledge. We hope to contribute some important steps towards handling this problem by making explicit a number of common sense oriented modelling decisions within the LILOG context. It is obvious, though, that both background knowledge for natural language processing and the adequate implementation of metainformation for knowledge base contents will be an ongoing affair for the next years.

Acknowledgement: We thank Bart Geurts, Tibor Kiss, Ewald Lang, Kai von Luck and Martin Mezger for useful ideas and stimulating discussions.

References

[Allen, 1983] James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

[Brachman and Schmolse, 1985] Ronald J. Brachman and James G. Schmolse. An overview of the KL-ONE knowledge representation system. 9(2):171–216, April 1985.

[Brachman et al., 1985] Ronald J. Brachman, Victoria Pigman Gilbert, and Hector J. Levesque. An essential hybrid reasoning system: knowledge and symbol level accounts in KRYPTON. pages 532–539, August 1985.

[Hayes, 1979] Patrick J. Hayes. The naive physics manifesto. In D. Michie, editor, *Expert Systems in the Microelectronic Age*, Edinburgh Univ. Press, 1979.

[Hobbs et al., 1987] Jerry R. Hobbs, William Croft, and Todd Davies. Commonsense metaphysics and lexical semantics. *Computational Linguistics*, 13(3-4):241–250, August 1987.

[Horacek, 1989] Helmut Horacek. Towards principles of ontology. In D. Metzing, editor, *Proc. GWAI-89*, pages 323–330, Springer, Berlin, Germany, 1989.

[Lehnert, 1988] W.G. Lehnert. Knowledge based natural language understanding. In H. Strobe, editor, *Exploring Artificial Intelligence*, pages 83–131, Morgan Kaufmann, San Mateo, 1988.

[Mann et al., 1985] William C. Mann, Yigal Arens, Christian M. I. M. Matthiessen, Shari Naberschnig, and Norman K. Sondheimer. Janus abstraction structure—draft 2. Draft paper, University of Southern California, Information Science Institute, Marina del Rey, Cal., October 1985.

[Monarch and Nirenburg, 1987] I. Monarch and S. Nirenburg. The role of ontology for knowledge-based systems. In B. Gaines J. Boose, T. Addis, editor, *Proc. EKAW-87*, Reading University, Reading, Mass., 1987.

[Palmer, 1978] Stephen E. Palmer. Fundamental aspects of cognitive representation. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorisation*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978.

[Pirlein, 1990] Thomas Pirlein. *Rekonstruktion von Hintergrundwissen für ein wissensbasiertes textverstehendes System.* IBM Deutschland GmbH, September 1990.

[Pletat and von Luck, 1989] Udo Pletat and Kai von Luck. Knowledge Representation in LILOG. In Karl-Hans Bläsius, Uli Hedtstück, and Claus Rollinger, editors, *Sorts and Types in Artificial Intelligence*, 1989.

[Tamas, 1986] G. Tamas. *The Logic of Categories.* W. H. Freeman and Company, Stuttgart, 1986.

[Trabasso and Sperry, 1985] T. Trabasso and L.L. Sperry. Causal relatedness and importance of story events. *Journal of Memory and Language*, (0):595–611, 24 1985.

[Vendler, 1967] Zeno Vendler. *Linguistics in Philosophy.* Cornell University Press, Ithaca, N. Y., 1967.

[14] "Das Museum wird um 11 Uhr geöffnet"

[15] "Das Museum ist von 9 bis 15 Uhr geöffnet"