Representation of Semantic Knowledge with Term Subsumption Languages

Gerrit Burkert, Peter Forster

Institut für Informatik, Universität Stuttgart
Breitwiesenstr. 20/22
D-7000 Stuttgart 80, Germany
e-mail:
burkert@informatik.uni-stuttgart.de
forster@informatik.uni-stuttgart.de

Abstract

One problem in the design of a lexicon for natural language processing is the representation of semantic knowledge. We examine the adequacy of knowledge representation formalisms developed in artificial intelligence, in particular of term subsumption languages for these issues. In order to derive some basic requirements for a suitable representation language we analyze a number of definitions of a monolingual dictionary.

1 Introduction

The lexicon as one of the major linguisitic knowledge sources of a natural language processing (NLP) system contains among others morphological, syntactic and semantic knowledge. The design of an appropriate lexicon involves the following questions:

- Which size is required for the lexicon for a given application?
- What are the units of the vocabulary?
- How should lexical information be represented?
- What kinds of techniques should be applied for generating a large lexicon, e.g., could the lexicon be partially generated by extracting information from machine-readable dictionaries?

In this paper we address the third problem, more precisely the *representation of semantic information.*

Semantic knowledge is crucial for both understanding and generating natural language. One important class of representation models for semantic knowledge are the network models, which have been influenced by early association models. In network models, word meanings can be described by relationships to other word meanings. In contrast, other models for representing semantic knowledge aim at the representation of structural aspects of word meanings so that a particular word meaning is built up by a set of particular semantic features.

During the last years, a number of knowledge representation (KR) systems have been built using term subsumption languages (TSLs). TSLs are formal languages for defining concepts by reference to superconcepts and by specification of additional features. Because of the similarity of concept descriptions in TSLs and dictionary definitions based on "genus proximum et differentia specifica", it seems useful to investigate to what extent TSLs can be used for the representation of word meanings.

In the next section, we discuss some properties of TSLs and their relation to other KR formalisms. In order to get some hints whether TSLs are suitable for the representation of word meanings we examine several dictionary definitions. Next, we give a short introduction to our terminological formalism and outline, how the formalism can be integrated into a lexicon. Finally, we summarize the results and give a short survey of our current work.

2 Knowledge representation and TSLs

In the field of AI, many researchers have addressed the problem of knowledge representation; in this area semantic networks played an important role. In semantic networks the knowledge is described by nodes and links. While Quillian aimed at the representation of word meanings [Quillian 68], semantic networks have also been used to model propositions, events, spatial relationships and so on. Since semantic networks failed in providing a unique semantic interpretation, several researchers examined the "semantics of semantic networks" ([Woods 75], [Brachman 79]).

Another approach is to organize knowledge in chunks called frames [Minsky 75] which are used to represent "stereotypical situations". Frames typically allow the specification of default slot values, perspectives and attached procedures. Collections of frames can be combined to frame-systems. The expressive power of frame systems makes it impossible to provide a well defined semantics for them.

Both, elements of different network formalisms and basics of the frame theory, have influenced the structural inheritance networks and the subsequent implementations (KL-ONE, [BrachmanSchmolze 85]). The basic idea is to postulate a level of knowledge representation with "knowledge structuring primitives, rather than particular knowledge primitives" ([Brachman 79]), the so-called *epistemological level*. The basic buildung blocks of KL-ONE representations are "concepts", i.e. structured conceptual objects. "Roles" are possible relationships between two concepts. The subsumption relation organizes the concepts in a concept taxonomy. Concepts are described with respect to their superconcepts by restricting and differentiating roles. In particular, roles can be restricted by the number (number restriction) and the range (value restriction) of allowed role fillers. If the specified restrictions constitute necessary and sufficient conditions for the concept, it is called a defined concept, wheras primitive concepts only need necessary conditions. Classification, an important inference mechanism of KL-ONE like systems, inserts concepts at the correct place in the concept hierarchy.

A logical reconstruction of KL-ONE revealed that the semantic status of a number of notions of KL-ONE was rather unclear. TSLs are formal knowledge representation languages derived from KL-ONE providing well-defined semantics which enables the decision whether the inferences are sound and complete. A number of KR systems based on TSLs have been developed, for instance, Krypton [Brachman et al. 85], KL-Two [Vilain 85], Back [Peltason et al. 89], Loom [MacGregorBates 87]. Besides a component for defining concepts and reasoning about the relationships between concepts (terminological component, TBox) these systems include an assertional component (ABox) that allows the definition of assertions about individuals.

3 Representing word meanings with TSLs

While aspects of syntactic structure are rather well understood in NLP, the problem of representing semantic information is far from being solved. Scientists from various research areas, e.g., linguistics, philosophy of language, lexicology and artificial intelligence, are dealing with problems concering the nature of word meanings and means for their representation.

In the following, we make some general remarks on semantic description without going into details of any semantic theory. A particular aspect all semantic theories are concerned with is the principle of compositionality: the meaning of a sentence is a function of the meaning of each of its components and its context. As a first approximation, one could abstract away from the context or assume a typical context (paradigmatic analysis). A good semantic theory, however, must allow the notion of semantic variation. So, in addition to the enumeration of possibly different senses of a word, we have to examine the meaning of a word in varying contexts (syntagmatic analysis). Also, a finite enumeration of word senses does not suffice to explain the creative use of words ([BoguraevPustejovsky 90]).

In addition to syntactic and semantic knowledge, the process of understanding natural language involves extralinguistic (encyclopedic) knowledge. Mechanisms for the combination of these types of knowledge are an important prerequisite for natural language understanding.

The analysis of word meanings is also the subject of dictionary definitions. During the last years, there has been an increasing interest in methods for extracting lexical semantic information from machine-readable dictionaries (see for example [BoguraevBriscoe 89]). There is, however, no consensus about the representation formalism into which the meaning descriptions should be transformed. In our opinion, the suitability of AI-based KR formalisms for the representation of semantic knowledge and as a means for the combination of linguistic and extralinguistic knowledge has to be investigated. As a first step, we are analyzing a number of simple dictionary definitions in order to derive some basic requirements which a representation language has to meet to be usable for the representation of word meanings. The results will allow us to assess the suitability of TSLs for that matter.

In a dictionary, different meanings of a word are usually specified by means of definitions, examples, references and pictures.[1] Subsequently, we will concentrate on the analysis of meaning definitions. There exist different types of definitions, e.g.,[2]

- definition by reference to synonyms
  - acclaim: applause; approval.
  - complaint: illness; disease.
  - jowl: jaw.
- definition by reference to antonyms
  - absolute: not relative.

[1] A comparison between different types of dictionaries and a closer investigation of the definitions in a dictionary can not be given here. See for example [Hausmann 85].

[2] Even though we investigated german definitions from the *Duden Bedeutungswörterbuch* ([Duden 85]) the definitions we mention in this section are taken from the machine-readable version of the *Oxford Advanced Learner's Dictionary* ([OALD 88]) because it turned out to be difficult to translate the german definitions without loss of information.

- affected: not natural or genuine.
- wild: (of plants) not cultivated.

- definiton by reference to hyperonyms and modifying elements
  - park: public garden or public recreation ground in a town.
  - bobsled: large, long sleigh with brake and steering wheel, used for racing.
  - blackboard: board used in schools for writing and drawing on with chalk.

We will have a closer look at nominal definitions of the latter type which contain a genus term of the defined word.[3]

The first part of the definition of park, namely "public garden", can be represented by a concept with superconcept **garden** and a relation called **PROPERTY** to the concept **public**:

a **park** is a
  garden
  with **PROPERTY** **public**

A visualization of this definition (in a KL-ONE-like graphical notation) is given in Fig. 1.

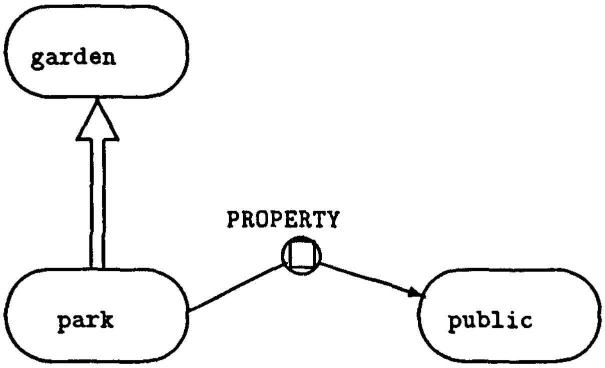

Figure 1: Representation of **park**

The concept identifiers, e.g., **garden**, in the example have to be distinguished from the corresponding word forms.[4] Each concept represents one of the possible meanings of the corresponding word.

This example shows at least two problems of representing word meanings with TSLs. The main problem arises from the fact, that the epistemological primitives of a TSL do not give enough specifications for the representation of word meanings. Many nominal definitions contain nouns modified with adjectives. We need a number of predefined

[3] In the German monolingual defining dictionary [Duden 85] around 70% out of a sample of 126 randomly chosen definitions of nouns are of that type.

[4] In order to distinguish between concepts and words concepts are in the following printed in **typewriter style**.

and modifiable roles, like the relation PROPERTY between the nominal concepts and the concepts representing these adjectives.

Another important relation for the representation of noun meanings is the part-whole relation, called meronymy. An example is the definition of *bobsled: large, long sleigh with brake and steering wheel, used for racing.* The concept bobsled refers to the parts brake and steering-wheel:

```
a bobsled is a
          sleigh
          with PROPERTY large
          with PROPERTY long
          with PART brake
          with PART steering-wheel
          with FUNCTION racing
```

This example shows a third class of important relations for the definition of noun meanings, i.e. the normal uses or functions of a thing. The relations can be further specialized, e.g., the PART relation describes different types of meronymy like COMPONENT, MEMBER, MATERIAL (see [Miller et al. 90]).

The part-whole relation is a relation between nominal concepts and can be represented in a TSL-based KR formalism by means of roles. Number and type of given parts can be described by number restriction and value restriction respectively. Different parts of a thing have to be specified by different subroles of a more general PART role. In the example above the two components brake and steering-wheel have to be related to bobsled by two COMPONENT roles:

```
a bobsled is a
          sleigh
          with COMPONENT1 brake
          with COMPONENT2 steering-wheel
          ...
```

The roles are organized in a role hierarchy:

```
COMPONENT is a PART
COMPONENT1 is a COMPONENT
COMPONENT2 is a COMPONENT
```

The PROPERTY relation is a relation between nominal concepts and "property concepts", e.g., public in the first example. Such kinds of concepts do not fit into a term hierarchy because they usually do not have suitable superconcepts or individuals. Consequently, the most important inference mechanism of TSLs, namely classification is unsuitable for the representation of property concepts. We presumably need another formalism for the representation of properties, in which other relations, for example antonymy, play an important role ([GrossMiller 90]). This formalism has to be combined with the term subsumption formalism.

The FUNCTION relation relates nominal concepts to concepts representing verb meanings, e.g.,

```
a bobsled is a
          sleigh
          with ...
          with FUNCTION racing
```

The representation of verb meanings involves a number of further problems, e.g., the representation of space and time, that can not be investigated in this paper.

Returning to the representation of nominal concepts, we try to represent the complete definition of *park: public garden or public recreation ground in a town*:

```
a park is a
      garden
      with PROPERTY public
  or is a
      recreation-ground
      with PROPERTY public
      with LOCATION town
```

This example demonstrates the necessity of concept disjunction. Disjunction is frequently used in definitions. Therefore, a KR formalism adequate for the representation of semantic knowledge has to provide a form of concept disjunction. Disjunction is not allowed in TSLs because it is contrary to the claim that concepts should only be defined with respect to their superconcepts. In the example, both "garden with property public" and "recreation-ground with property public and with location town" (subsequently termed **p-garden** and **p-recreation-ground** respectively) are subconcepts of **park**, as illustrated in Fig. 2.

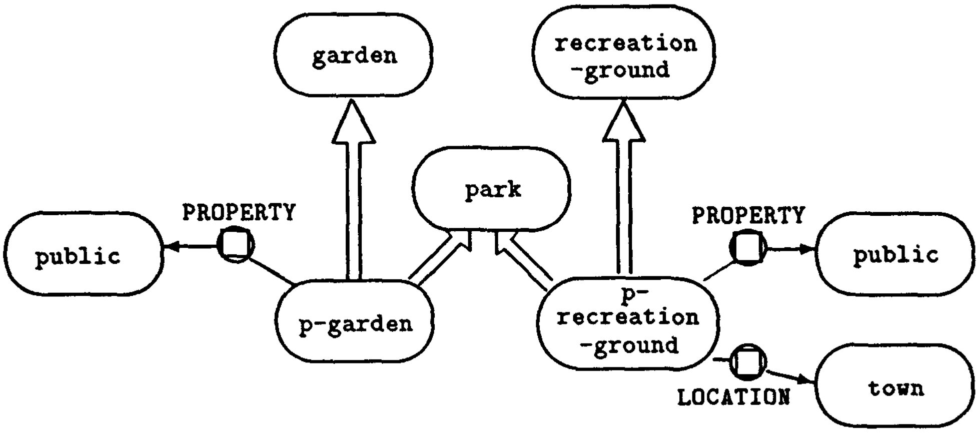

Figure 2: Another representation of **park**

This representation of **park** is unsuitable because it does not guarantee that all parks are public gardens or public recreation grounds in a town. A solution to this problem is the notion of a "covering", which was inroduced in NIKL ([Moser 83]). If **park** is defined as a covering of **p-garden** and **p-recreation-ground**, every instance of **park** will be an instance of at least **p-garden** or **p-recreation-ground**. In NIKL coverings are used to enhance concept specifications but they are ignored by the classifier.

The investigation of further nominal definitions revealed that several other extensions to TSLs are necessary:

- boil: hard (*usu* red, *often* painful) poisoned swelling under the skin, which bursts when ripe.
- boulevard: wide city street, *often* with trees on each side.

These examples indicate the existence of typical features of a concept, called "defaults". Inferences with defaults require the use of nonmonotonic reasoning techniques which are outside the scope of a TSL classifier.

The following examples show that similarity between concepts is another important relation.

- lemur: nocturnal animal of Madagascar, *similar to* a monkey but with a foxlike face.
- marimba: musical instrument *similar to* the xylophone.
- quail: small bird, *similar to* a partridge, valued as food.

Most of the requirements mentioned above can not be integrated into a TSL maintaining sound and complete inferences. Because these requirements seem to be necessary for the representation of word meanings, TSLs can provide only a "representational kernel", which has to be embedded into a component with greater expressive power. This component has to allow enhanced concept descriptions, e.g., concept disjunction, defaults and similarity of concepts.

4 An Approach to the Integration

As a first step we have implemented a KR system that consists of a terminological and an assertional component. The formalism used is similar to the formalism described in [Nebel 90]. The restricted expressiveness enables inferences that are sound and complete and makes the formalism suitable as a platform for the extensions described above.

A small fragment of the TBox language is illustrated in the following example. The concept `bobsled` is described as a subconcept of `sleigh` with two `PART` relations, namely `COMPONENT1` and `COMPONENT2`. These roles have to defined separately as specializations of `COMPONENT`. The concept `bobsled` is primitive because the specifications are necessary but not sufficient for the definition of bobsled.

```
(defrole (PART))
(defrole COMPONENT (PART))
(defrole COMPONENT1 (COMPONENT))
(defrole COMPONENT2 (COMPONENT))

(defconcept bobsled (sleigh) :primitive
            (:all COMPONENT1 brake)
            (:all COMPONENT2 steering-wheel))
```

The system is implemented in CLOS (Common Lisp Object System) and an overview of its syntax and semantics is given in [Forster et al. 91].

The formalism has to be integrated into a lexicon, a possible architecture of which is shown in Fig. 3. The lexicon consists basically of two components: one containing word forms and another for the representation of word meanings. The latter has to be embedded into a component for the representation of more general world knowledge.

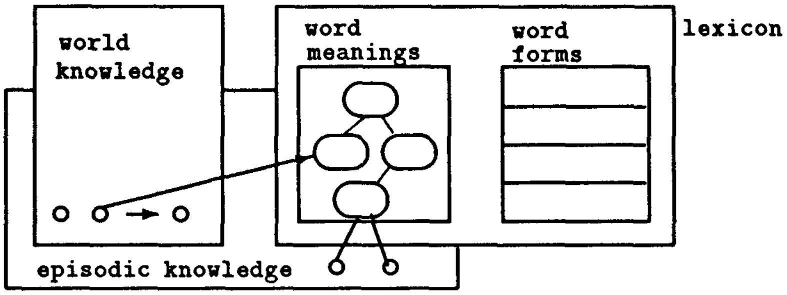

Figure 3: Integration of the components

5 Conclusion

We have outlined that TSLs are KR formalisms with some interesting features. The well-defined semantics allows to decide whether the inference algorithms are sound and complete.

In order to use TSLs for the representation of word meanings we examined a number of dictionary definitions. TSLs seem to be adequate for the representation of some aspects of noun meanings, for example the subsumption and part-whole relations. There are, however, important aspects of word meanings which can not be expressed in TSLs, e.g., concept disjunction, defaults and similarity between concepts. As a consequence the TSL formalism has to be embedded into a representation system with more expressive power.

References

[BoguraevBriscoe 89] B. Boguraev and T. Briscoe (Eds.). *Computational Lexicography for Natural Language Processing.* Longman Group UK Limited, 1989.

[BoguraevPustejovsky 90] B. Boguraev and J. Pustejovsky. Lexical Ambiguity and Knowledge Representation. In N. V. Findler (Ed.), *13th International Conference on Computational Linguistics*, Helsinki, 1990.

[Brachman et al. 85] R. Brachman, V. Gilbert and H. Levesque. An Essential Hybrid Reasoning System Knowledge and Symbol Level Accounts of KRYPTON. In *Proc. of IJCAI-85*, pp. 532–539, Los Angeles, CA, August 1985. IJCAI, Inc.

[Brachman 79] R. Brachman. On the Epistemological Status of Semantic Networks. In N. V. Findler (Ed.), *Associative Networks - The Representation and Use of Knowl-*

edge by Computers. Academic Press, New York, 1979. Also BBN Report 3807, April 1978.

[BrachmanSchmolze 85] R. J. Brachman and J. Schmolze. An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, 2(9):45, 1985.

[Duden 85] Duden. *Duden Bedeutungswörterbuch*. Bibliographisches Institut, 1985.

[Forster et al. 91] P. Forster, G. Burkert and O. Eck. Wissensrepräsentation mit TED und ALAN. Forthcoming: Interner Arbeitsbericht, Institut für Informatik, Universität Stuttgart, 1991.

[GrossMiller 90] D. Gross and K. J. Miller. Adjectives in WordNet. *International Journal of Lexicography*, 3:265–277, 1990.

[Hausmann 85] F. J. Hausmann. *Lexikologie*, pp. 367–411. Athenaum Verlag, 1985.

[MacGregorBates 87] R. MacGregor and R. Bates. The Loom Knowledge Representation Language. Technical Report, USC/ Information Sciences Institute, 1987.

[Miller et al. 90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244, 1990.

[Minsky 75] M. Minsky. A Framework for Representing Knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision*, pp. 211–277. McGraw-Hill, New York, 1975.

[Moser 83] M. Moser. An Overview of NIKL, the New Implementation of KL-ONE. In *Research in Natural Language Understanding*. Bolt, Beranek, and Newman, Inc., Cambridge, MA, 1983. BBN Technical Report 5421.

[Nebel 90] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems*. Lecture Notes in Artificial Intelligence. Springer Verlag, 1990.

[OALD 88] OALD. *Oxford Advanced Learner's Dictionary Electronic Version*. Oxford University Press, Cambridge, 3rd edition, 1988.

[Peltason et al. 89] C. Peltason, A. Schmiedel, C. Kindermann and J. Quantz. The BACK System Revisited. Technical Report kit - report 75, Projektgruppe KIT - Fachbereich Informatik- TU Berlin, 1989.

[Quillian 68] M. Quillian. Semantic Memory. In M. Minsky (Ed.), *Semantic Information Processing*. The MIT Press, Cambridge, MA, 1968. Also PhD Thesis, Carnegie Institute of Technology, 1967.

[Vilain 85] M. Vilain. The restricted language architecture of a hybrid representation system. In *Proceedings of IJCAI-85, Los Angeles, Ca.*, pp. 547–551. IJCAI, 1985.

[Woods 75] W. Woods. What's in a link: Foundations for semantic networks. In D. G. Bobrow and A. Collins (Eds.), *Representation and Understanding: Studies in Cognitive Science*, pp. 35–82. Academic Press, New York, 1975.