Hennie van der Vliet
Free University of Amsterdam

Conceptual Semantics for Nouns

Abstract

This contribution deals with the development of a conceptual lexicon for NLP–purposes. In this article our focus will be on nominals. The conceptual lexicon contains commonsense knowledge that is shared by language users. In the lexicon each word is related to a concept that is defined by a set of relations with other concepts. The (multiple) inheritance of additional relations is governed by a concept–hierarchy. The acquisition of knowledge is arranged by semi–automatically parsing a dictionary, using a tagger–lemmatizer, a linguistic knowledge bank and a dictionary of synonyms. However, the parsing process is not restricted by the use of any specific knowledge source, any tagger–lemmatizer in combination with information on selection restrictions will do. The result will be a multi–purpose lexicon which can be used for instance in on–line dictionaries and in an NLP environment, improving the extend to which NLP systems can cope with incomplete information and can solve ambiguity.

1. Introduction

The lexicology research team at the Free University of Amsterdam has built a lexical databank for the processing of natural language. This databank, called SNIV (Subcategorisatie Nederlands IBM–VU), already contains a great deal of graphemic, morphologic, syntactic, semantic and pragmatic knowledge on nouns, verbs and adjectives. In order to add commonsense knowledge to SNIV, we are working on a module for conceptual semantics. We believe that conceptual lexicons containing this kind of encyclopedic knowledge may be used to provide useful representations of what phrases and sentences are actually about, by explicating relevant parts of the knowledge and beliefs underlying the words. Moreover, the lexicon will improve the overall performance of NLP systems in coping with incomplete information and in solving ambiguity.

Unlike most studies we will be focussing on the conceptual semantics of nouns. Research in conceptual semantics has primarily focussed on verbs, because the semantics of verbs are important from a compositional point of view: they provide the argument structure of a sentence. Our approach to semantics is lexical rather than compositional.[1] It will be pursued below that the categories to which the nouns refer, contribute significantly to the actual meaning of language. Moreover, it will be argued that conceptual meaning is a very important, if not the cardinal meaning aspect in the semantics of nouns.

This paper provides a brief overview of the project. In Section 2 our approach to conceptual semantics is presented. In Section 3 we comment on the representation of conceptual semantics by a relational database in combination with a concept hierarchy. The semi–automatic acquisition of conceptual knowledge, which is the subject of the next section, is based on semantic strategies. We use frames as patterns of expectation for extracting knowledge from machine–readable dictionaries. The process of acquisition is only semi–automatic, because the frames are drawn up manually. In Section 5 we discuss the resulting conceptual lexicon and we suggest possibilities for further research, with regard to the evaluation of the lexicon and the acquisition of additional knowledge gained when using the lexicon in NLP sessions.

2. Conceptual semantics

In our view the semantics of a word may be defined as a combination of different types of meaning, including for instance conceptual, grammatical, collocational and associative meaning (see Martin 1992). The different aspects of meaning are not equally represented in each word; i.e. one aspect tends to be predominant. In this section we will discuss conceptual meaning and its representation. First we will discuss the relevance of conceptual meaning for nouns.

A concept can be extensionally defined as a set of objects. However, there is little room for lexical semantics in this view. To get a grip on the semantics of words, an intensional definition is necessary. From an intensional point of view a concept may be described as a list of conditions governing the recognition of objects as members of categories. Research in cognitive semantics indicates that, as a rule, these conditions are neither necessary nor sufficient, they function as defaults or patterns of expectation (e.g. Taylor 1989). The categories are based on graded membership: the objects may be more or less central to the category.

The SNIV–module for conceptual semantics is based on this cognitive point–of–view. In our terms, a concept is the representation of the knowledge which members of a linguistic community use in categorizing their world. Since the lexicon reflects this categorization, the concepts are the link between commonsense knowledge and the lexicon. The concept 'cat', for instance, is a representation of what people know about cats: they are furry pets, they catch mice and birds and they drink milk. As such the concepts are chunks of knowledge, linked to the lexicon, but not lexical in themselves.

A conceptual lexicon, which is to be used in an NLP environment, should contain linguistically relevant knowledge which is shared by the users of a language. Obviously it is impossible to collect all commonsense knowledge in a database, let alone to determine the extent of linguistic relevance of all this knowledge. Apart from the problem of acquisition and evaluation, there is the problem of computational efficiency; after all the module should be working in an NLP environment. Finally, in agreement with the cognitive theory, we are convinced that once the central body of knowledge is obtained, the acquisition of lots of peripheral features will only marginally incresase the qualitative results. In practice this means that we do not tax ourselves unduly by trying to find the most extensive conceptual representation for each word. Instead we merely try to find the central properties, the properties that are most important in linguistic categorization, ensuring that the representations of the concepts can easily be changed and extended.

As noted above, conceptual meaning is just one aspect of lexical meaning. Not all the aspects of meaning are equally represented in different words. In the case of prepositions, for instance, the functional–grammatical meaning is of cardinal importance. Domain specific terms are the classic example of rich conceptual meaning (Martin 1992:3). Although it is clear that not all nouns share this property, conceptual meaning may be considered to be their cardinal aspect of meaning. This applies specifically to concrete nouns. A word like *car* reveals a huge amount of knowledge about vehicles in general, specific cars, their properties and parts, their influence on the environment etc. This knowledge is linguistically relevant, because many sentences about cars cannot be properly understood without it. The next sentence for example, taken from the definition of *car* in the *Collins Cobuild English Language Dictionary*, presupposes that cars are driven by human beings:

(1) The car drove off, and Mrs. Foster was left alone.

This wealth of conceptual meaning and its relevance to our understanding of language, motivated our decision to concentrate on nominals rather than verbs.

In conclusion, a conceptual representation of commonsense knowledge provides a description of part of the semantics of words. It is relevant for natural language processing in that it allows us to express part of the meaning of a sentence; not with regard to truth values, but in terms of knowledge and beliefs. What is perhaps even more important is that a conceptual lexicon may also serve as a reference point for interpretation. As a reference point it can deal with incomplete information and ambiguity in a very natural and efficient way, by providing expectations that function as guidelines in syntactic and semantic analysis.

3. Conceptual semantics in SNIV

The conceptual lexicon is subdivided in a set of categories. Each category is a hierarchically organized network of members: *pet* is part of the category *animal*, *dog* is a pet and *pitbull* is a dog. There is no formal difference between roots and terminal nodes, they are all concepts; the top and the bottom of a hierarchy are arbitrarily fixed.

A category is defined by a set of potential properties that are shared by its members: pets may be animate, they may be owned by someone, they may be part of a family and they may live in a house. These properties are potential because they are merely expectations. They are not necessary or sufficient for membership of a category, but as a set they do function as a criterion of graded membership. The properties which distinguish concepts are represented in relations. Each relation is a three-place predicate. The first predicate is the name of the defined concept, the second is the relation and the third is the related concept:

(2) rel(cat,has_part,whiskers).
rel(cat,function,companionable).

The members of the pet-concept inherit the relations from *pet* and from *animal* (the root). The fillers for the related concepts are inherited by default, but they may be overruled, specified or blocked. The inheritance of the values of hyperonym concepts is guided by the hierarchical structure of the categories. The categories are represented as networks of isa-relations:

(3) rel(pet,isa,animal).
rel(feline,isa,animal).
rel(cat,isa,pet).
rel(cat,isa,feline).

As was pointed out above, there is not just one, but a set of networks, since there is no ultimate root or 'superconcept' that covers all other concepts. Each concept may be a member of several networks and may therefore inherit relations from each of these (multiple inheritance). As a consequence the concept *cat* may inherit properties both from *pet* and from *feline*. Although multiple inheritance can give rise to problems such as contradictory relations, it is a very powerful tool for storing information in an economical way (see Touretzky 1987).

In related research the set of properties defining a category is often referred to as a frame (e.g. Martin 1992). We only use this term in the context of the acquisition of the relations and their values, since once they are gathered, they are stored separately in the database. There is no need to organize them in frames, because the frames can be produced within the database: the frame for *cat* is the subset of the relations with *cat* as a first argument. In the same way the subset of all pets can be found, the subset of all the parts of pets, the subset of pets with wings and even the subset of pets with wings that are not birds. In conclusion, all kinds of operations on sets are possible, because the lexicon functions both as a database and as a flexible inference system.

4. Acquisition

Building a semantic network for everyday knowledge is an ambitious project. We already mentioned the storage problem, but an even greater problem is the acquisition. In principle, all human knowledge can be relevant, but we are specially interested in the most characteristic knowledge about words, the kind of knowledge one expects to find in a dictionary. Accordingly we decided to carry out the acquisition by conceptually parsing dictionary definitions.
We make use of four sources:

- a machine readable version of the *Van Dale Basiswoordenboek Nederlands* (BVD)
- a machine readable version of the *Van Dale Groot Woordenboek van Synoniemen* (SVD)
- the SNIV databank
- D–TALE, a tagger–lemmatizer for Dutch, developed by the lexicology research team at the Free University of Amsterdam

The BVD is a dictionary for children between the ages of 10 and 15. It contains 25000 frequently used Dutch words. The style and content of the definitions are simple and straightforward, which makes them very suitable for automatic knowledge extraction.

The SVD contains about 45000 entries, subdivided into 2000 categories. The machine readable version can be seen as a set of monotonic hierarchical networks, in which each network represents a category. An example of such a category is *vervoermiddel* (means of transport). The path in the network for *taxi* (*taxi*) looks like this:

(4) vervoermiddel (means of transport)
  —> voertuig (vehicle)
    —> motorvoertuig (motor vehicle)
      —> auto (car)
        —> taxi (taxi)

By transforming the SVD networks into isa–relations, we are building up the hierarchical networks mentioned above. The information in the SVD is also used to conceive a flexible way of pattern matching: if the conceptual parser expects *vehicle*, but finds *car*, the SVD causes it to conclude that there is a perfect match. In other words, the related concept may be specified by any concept it dominates.

The output of the tagger–lemmatizer D–TALE is the basis for an elementary syntactic preprocessing. Out of the tagged and lemmatized definition strings the preprocessor builds minimal np's and pp's, and determines their heads. Nominal heads are important because they are candidates to be the object of a relation. Prepositions provide extra evidence for the kind of relation that is involved.

The SNIV–databank provides the syntactic information to rally round the preprocessor. Moreover it provides markers for selection restrictions on nominals, like <animate>, <substance> and <artefact>. These markers are used for the acquisition of conceptual knowledge, as will be discussed below.

Our means of knowledge acquisition through conceptual parsing is not merely restricted to the use of the dictionaries we mentioned above. Most dictionaries will meet the needs of conceptual parsing, as long as they provide the everyday encyclopedic knowledge that conceptual semantics is based on. The same holds for the tagger–lemmatizer and the hierarchical network. Although the acquisition is not restricted to specific dictionaries and tools, it is obvious that the quality of the knowledge resources will have a serious impact on the results.

Having discussed the resources of knowledge we use, we now shift to the knowledge acquisition. The first step in the acquisition is building a frame for a specific category[2]. This is done manually, based on words that are considered to be prototypical members of the category, because they are expected to contribute the most relevant relations. The categories and prototypes are provided by the SVD[3]. Some examples of relations for **vervoermiddel** are **has_part**, with an artefact as a related concept, and **transports** with human beings (as passengers) or concrete objects (as freight) as a related concept. The relations are presented in three–placed predicates, the first argument being the subject of the relation, the second the kind of relation and the third the object of the relation:

(5) rel(vervoermiddel,has_part,artefact).
rel(vervoermiddel,transports,passenger).
rel(vervoermiddel,transports,freight).

The relevant definitions are tagged and lemmatized by D–tale and the results are interpreted by the preprocessor: the nominal heads are sorted out, adjectives are attached to the nominals and the prepositions are related to the nominal heads of the np's which are imbedded in the pp's. As an example the entry for *aak (*barge*)* in the BVD is presented:

(6) *aak*
boot met een platte bodem voor vrachtvervoer over rivieren en kanalen
(*barge*
ship with a flat bottom for goods carriage on rivers and canals)

D–tale provides a list of the tagged and lemmatized words in the definition and the preprocessor builds the minimal np's and pp's. As a results the nominal heads *boot*, *bodem*, *vrachtvervoer*, *rivier* en *kanaal* are proposed to be in some relation to *aak*. Since *bodem* is in a prepositional phrase, its relation to *aak* is expressed by *met*. The relations between *aak* and *rivier* and *aak* and *kanaal* are expressed by *voor*. In addition the adjective *plat* (the citation form of *platte*) is attached to *bodem*: the property 'plat' holds for *bodem* if *bodem* is in relation to *aak*. This information is represented in the same format as the frames:

(7) rel(aak,REL,boot).
rel(aak,REL,bodem).
rel(aak,prep(met),bodem).
rel(aak,prop(plat),bodem).
rel(aak,REL,vrachtvervoer).
rel(aak,prep(voor),vrachtvervoer).
rel(aak,REL,rivier).
rel(aak,REL,kanaal).
rel(aak,prep(over),rivier).
rel(aak,prep(over),kanaal).

After the syntactic preprocessing of the definitions, the proposed relations are compared with the relations in the frames to select the relevant second argument, the kind of relation. Relevant information is provided by the trees of the SVD, the entries of the SNIV databank and can be found in the preposition–relations. The values of the third argument of the frame–slots serve as selection restrictions for the nominal heads in the third argument of the proposed predicate: if some third argument in a proposed vehicle–predicate according to SNIV is an artefact, it may be combined with the has_part relation (recall that SNIV provides all the relevant information by the markers). In addition to the selection restrictions the process is guided by prepositions. Prepositional heads provide extra evidence for the filling of the slots by the nominal heads of embedded np's. This is necessary to prevent the system from errors and because several slots in a frame might be specified for the same selection restriction.

As an example I again use the *aak*–predicates. For each of these *aak*–relations with a variable as a second object the question is which of the frame–relation is involved. In the SVD *aak* is called a *vervoermiddel* and as a result the *vervoermiddel*–frame is activated. In the case of *rel(aak,REL,bodem)*, SNIV gives the information that *bodem* in one of its meanings is an artefact. As is shown in 5), artefacts may serve as the object in a has_part–relation. The preposition *met* is an additional indication and as a result the *rel(voertuig, has_part,artefact)* and *rel(aak,REL,bodem)* can be merged, resulting in *rel(aak,has_part,bodem)*.

In the system sketched above, the knowledge acquisition may be seen as a conceptual analysis of a dictionary. The bootstrapping problem of conceptual parsing without a knowledge base is solved by manually constructing the frames providing a set of relations which function as expectation patterns of what may be found in the definition. In the next section the structure and properties of the resulting system are discussed.

5. The results

The outcome of the process of acquisition is a network and a database:

- a multiple inheritance network representing the conceptual hierarchy
- a set of relations representing the properties.

The hierarchical network of isa–relations is a multiple inheritance network because the hierarchy originates both from the SVD and the BVD. In the BVD for instance *caravan* is called a small cottage (on wheels), while according to the SVD a caravan is a vehicle. Both the hierarchies will provide useful information.

All relations, the isa–relations as well as the others, are small chunks of conceptual meaning of both the defined and the related concepts. The database–relations of each concept originate from the frames of the categories the concept belongs to and the value of the third arguments, the objects of the relations, are specified by the information in the dictionary. Not all objects are specified by a concept, they are minimally specified by a semantic marker. On the other hand there are relations originating from the dictionary in which the proper relation cannot be specified. In (8) the first relation is to be found in the frame, the second is a result of the parsing of the defiition, as is shown in (7).

(8) rel(aak,transports,{HUMAN/FREIGHT}).
rel(aak,REL,vrachtvervoer).

Until now the system is not able to recognize *vracht* (goods) in the compound and as a result both the object of the transport–relation and the relation of the *vrachtvervoer*–predicate cannot be detected. However, the database does indicate that a barge is expected to transport whatever may be called human or freight and that a barge has an unspecified relation with goods carriage.

Apart from these explicit relations and categories, there are many implicit relations and categories, that may be derived from a combination of the explicit relations and the inheritance network: if the transport–relation for *car* has a human object, all humans in the database will be appropriate passengers for *car* and all its hyponyms. This can be overruled by an explicite object. If for instance *driver* is the object of the used_by–slot, it cannot also occur as a passenger. Apart from this relations can be derived in an indirect way, because concepts are not only the subject of the relations in their frame but also be the object in any other frame. As a result, the knowledge of a concept in the lexicon is not restricted by the definition of the corresponding word in the BVD; the relations based on information provided by other definitions are also available.

As was pointed out in Section 2, the formalism of the relations supports the creation of subsets of relations by putting restrictions on the subject, on the relation, on the object and on combinations of these. By producing subsets, new categories may be defined, such as the category of artefacts that have something to do with eating a meal. Such a subset is created by searching for all the relations with *meal* or a hyperonym of *meal* as a subject or as an object. Generation by inference is an illustration of the power of the semantic description as a result of the organizational set–up of the conceptual lexicon.

We would like to emphasize that the acquisition should be an ongoing proces, of which the parsing of a dictionary is only the beginning; after all, acquisition from dictionaries will only provide a very small selection of the everyday knowledge of language users. The objectives of further study will be a permanent evaluation of the relations, the slots and the fillers. This may for instance be done by keeping record of which of the slots are actually used in NLP–sessions. A related topic is additional knowledge acquisition, in which unknown words are stored and a record is kept of what (words belonging to) categories cannot be retrieved during NLP–sessions. In this way, the parsing of dictionaries may be seen as a way of bootstrapping.

6. Conclusion

The objective of the present article was to describe the structure of a conceptual lexicon we are constructing as a module in a lexicon which will be used for NLP–purposes, and to describe the acquisition of the conceptual knowledge.

The structure of this module for conceptual semantics is simple. The conceptual semantics of a word are described in terms of a set of relations to other words. Yet the structure provides for multiple inheritance and for the inference of implicit relations and categories. The acquisition of the conceptual knowledge can be described as a conceptual parsing of a dictionary. Since the sets of relations must be built manually, the process is only semi–automatic.

For the building of the system a dictionary, a tagger–lemmatizer and information on selection restrictions are needed. We also use a machine readable dictionary of synonyms, which is helpful. We would like to stress that any dictionary and tagger will do. In an NLP–environment, the module will be a powerful tool for the interpretation and the representation of the conceptual semantics of a sentence and may be of help in treating notorious NLP problems such as ambiguity and incomplete information. Other uses of the conceptual lexicon are for instance lexical disambiguation and term–generation in on–line dictionaries.

At this moment the system is only in an experimental stage and there is still research to be done on the development of the lexicon. In the last section we made some suggestions as to how the lexicon may be evaluated and developed while in use.

Notes

[1] This does not mean we deny the importance of compositional semantics. We merely believe that there is much more to semantics than compositionality alone. Semantics should not only provide the basic information for a semantical analysis of syntactic rules, it should also account for the meaning of words as a reflection of the way speakers use the language to represent the universe. After all language is a way to communicate ideas.

[2] In Section 3 we argued against the use of the term frame, because all relations are stored separately in a database. This is still true, but the relations are created in frames, since a category is defined by a set of relations. Once the relations are stored in the database, the frames are only retrievable by set–operations.

[3] Prototypes in the SVD can be recogized because they are often morphological simplexes, they have a lot of hyperonyms and they are not the root of a category (they should be 'basic level categories', according to Taylor 1989). Obviously this is not a scientific statement, but a practical way to set up a hypothesis.

References

Daelemans, W. and E.J. van der Linden 1992. *Evaluation of lexical representation formalisms: Tilburg,* ITK Research Memo

Dahlgren, K. 1988. *Naive semantics for natural language understanding.* Boston: Kluwer Acadamic Publishers.

Evens, M. (ed.) 1988. *Relational models for the lexicon: Representing knowledge in semantic networks.* Cambridge: Cambridge University Press.

Habel, C. 1985. "Das Lexikon in der Forschung der kunstmatigen Intelligenz" in C. Schwarze und D. Wunderlich (eds.), *Handbuch der Lexikologie.* Königstein: Athenäum.

Huijgen, M. and M. Verburg 1987. *Basiswoordenboek van de Nederlandse Taal.* Van Dale Lexicografie:Utrecht.

Jackendoff, R. 1990. *Semantic structures.* The MIT Press: Cambridge

Kayser, D. 1988. "What kind of thing is a concept?". *Computational Intelligence* 4:158 – 156.

Martin, W., H. Demeerseman en M. Vliegen 1992. *SNIV–project, beschrijving, opzet en verantwoording.* Amsterdam: VU.

Martin, W. 1992 "On the parsing of definitions" in: *Proceedings of the 5th Euralex International Congress.* Tampere.

Sterkenburg, P.G.J. Van e.a.1991 *Groot woordenboek van synoniemen en andere betekenisverwante woorden.* Utrecht/ Antwerpen:Van Dale Lexicografie

Taylor, J.R.1989 *Linguistic catgorization, prototypes in linguistic categories.* Oxford:Clarendon Press.

Touretzky, D.S. a.o.1987 "A clash of intuitions: The current state of nonmonotonic multiple inheritence systems." in: *Proceedings of the 10th international joint conference on artificial intelligence.* Milan.