Stefan Langer
Selectional classes and hyponymy in the lexicon

Doctoral thesis, CIS, University of Munich, 1995

This thesis describes the semantic encoding of nouns for the German electronic dictionary CISLEX. The semantic information is designed to serve a variety of purposes, which include the description of selectional constraints, the analysis of compound nouns and disambiguation of polysemic words.

Semantic classes and features of semantic classes

After the evaluation of existing encodings - e.g. the encoding in WordNet - I have designed an encoding formalism for the nouns in the CISLEX dictionary. The core of the semantic encoding for the nouns is the indication of one or several semantic classes for each lemma. A semantic class can be viewed as a node in a hyponymic structure of the vocabulary similar to WordNet, such as BIRD or PROFESSION.

The main difference between the approach described in the thesis and other approaches is the inclusion of information about different features of semantical classes in our encoding. Whereas often classes of lexemes obeying common selectional restrictions on the one hand and taxonomic classes on the other are conflated, and selectional restrictions and combinatoric features are used to define taxonomic classes, we make a difference between the features 'taxonomic' and 'selectional'. A pure class of the first type would certainly be one like BIRDS, whereas pure selectional classes, due to their definition,cannot be subsumed under a single expression, and have to be defined by means of typical contexts. The distinction of these two features allows a quick evaluation of the suitability of classes for different types of applications - whereas for text generation, machine translation etc. selectional classes are more relevant, for purposes of information retrieval only taxonomic classes have to be considered.

The initial assumption of a certain uniformity between selection and taxonomic organisation, underlying many other approaches to semantic encoding, is partly supported in the CISLEX-classification, as many classes - but not all - have both features 'selectional' and 'taxonomic'.

Classes in CISLEX are also marked for features indicating the internal structure ('uniform' versus 'prototypical') and for fuzzy versus exact borders.

A thesaurus structure of the lexicon is provided by a different kind of semantic class which is marked for the feature 'thematic'. Single lexical units and semantic classes as a whole can be linked to one or several thematic classes.

In addition to the hyponymic and thematical structure, the semantic encoding provides links to synonyms, meronyms and antonyms, and some minor semantic relations.

Coverage

CISLEX is a complete German dictionary. The semantic encoding should cover all nouns in the dictionary. We have adopted the following practical approach: All simple nouns (not prefixed and not complex forms), about 40 thousand lemmata have been manually encoded. The majority of about 10 thousand prefixed nouns in CISLEX has also been encoded. For the other complex forms, especially compound nouns, possibilities for an encoding based on the code for the simple forms have been investigated. This seems to be the only viable approach, as compounding is an extremely productive pattern in German word formation.

Compound nouns

In addition to the description of the encoding of simple nouns the thesis describes first attempts to find a strategy to encode German compound nouns. These experiments did also prove that the encoding of semantically non-complex forms is suitable for the statistical analysis of the semantics of compound nouns.

The number of compound lexemes occuring in corpora largely exceeds one million and new compounds are always formed, which makes it impossible to encode all occuring compound nouns manually. Using the classes encoded for the simple nouns, we have carried out studies on disambiguation of compounds. These first tests have shown that semantic patterns can be detected in large compound noun corpora, which could be exploited for an automatic or semi-automatic analysis, making use of a large, one million words corpus of compounds consisting of two nouns. Sense disambiguation of compounds and their parts corresponding to these classes includes the choice of the appropriate class for a lexeme in a text, when it has several semantic categories in the lexicon entry. It also includes the selection of a relation between the parts of the nouns, based on the semantic category of the parts. The probabilistic analysis of the corpus showed that it is possible to identify clusters of compounds where parts have similar relations to each other. Also nouns having polysemic parts can be classified according to the different meanings of the parts.

Semantic classes and text retrieval

In tests carried out on a sample of newpaper texts, it could be shown that the classes in CISLEX are also suitable as a basis for text retrieval. For this purpose, semantic classes and lexical units were bundled to build thematic classes. Subsequently, the texts were tagged with the thematic classes and then statistically classified.

Availability

The full thesis is only available in German language. The German title is:

Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX.
München: CIS-Bericht
ISBN: 3 - 930859-06-8

Download:

- ps-file (6MB!)
- zip compressed ps-file (1MB)

To order the printed version of the thesis contact the author:

stef@cis.uni-muenchen.de

Another publication (in English) which describes some work carried out for the thesis is:

Langer, Stefan, Petra **Maier** and Jürgen **Oesterle** (**1996**): CISLEX, an electronic dictionary for German. Its structure and a lexicographic application. Proceedings of COMPLEX 1996. Budapest.

CISLEX information

CIS homepage

Stefan Langer's homepage