THE ROLE OF SEMANTIC PROCESSING IN AN AUTOMATIC SPEECH UNDERSTANDING SYSTEM

Astrid Brietzmann and Ute Ehrlich
Lehrstuhl fuer Informatik 5 (Mustererkennung)
Universitaet Erlangen-Nuernberg
Martensstr. 3, 8520 Erlangen, F. R. Germany

Abstract We present the semantics component of a speech understanding and dialogue system that is developed at our institute. Due to pronunciation variabilities and vagueness of the word recognition process, semantics in a speech understanding system has to resolve additional problems. Its main task is not only to build up a representation structure for the meaning of an utterance, as in a system for written input, semantic knowledge is also employed to decide between alternative word hypotheses, to judge the plausibility of syntactic structures, and to guide the word recognition process by expectations resulting from partial analyses.

1. Introduction

Understanding spoken utterances requires more than mere word recognition. It is based on a number of meaning aspects, covering the range from textual interpretation of a sentence up to the revelation of the speaker's intention in the context of a special dialogue situation. In the speech understanding and dialogue system EVAR /4/, task-independent semantic analysis, domain-dependent pragmatic analysis, and dialogue-specific aspects are implemented in three separate modules /2/. Semantic analysis comprises those aspects that can be studied at the isolated sentence, independent from its actual use in the dialogue. The semantics module disregards communicative aspects of an utterance as well as its situational and thematic context. Thus, semantic consistency of words and constituents and underlying relational structure of the sentence are the main points of interest in this stage of analysis. Semantic knowledge consists of lexical meanings of words and selectional restrictions between them. The analysis of the functional structure is based on the principles of case and valency theory.

2. Valencies and case theory

The theoretical background for the analysis of functional relations in a sentence is given by *valency* and *case theory* /5, 3/. The main idea is that the syntactic and the semantic structure of a sentence are essentially determined by its head verb. The property to call for a certain number and kind of complementary noun groups or prepositional groups that are necessary to build up an adequate sentence is called *valency*. The morpho-syntactic and semantic descriptions of the complements constitute a verb frame with slots to be filled up by actual phrases. This valency frame is augmented by case labels circumscribing the functional role of the expected phrase. To give an example, the verb "suchen" (to look for) has the case slots:

AGENT: noun group (nominative), ANIMATE, obligatory
OBJECT: noun group (accusative), obligatory
LOCATION: adverbial group, PLACE, facultative.

The lexical knowledge base provides caseframe entries for all verbal and nominal items with valency properties. Mostly, meaning alternatives correspond to different caseframes. We use a relatively detailed case system with about 30 cases.

For use within the semantics module, a preprocessor transforms the dictionary entries to a network representation of concepts. The network scheme is influenced by the formalism of *Structured Inheritance Networks* /1/ and is described in /2/. It is used for knowledge representation in all semantic and pragmatic modules in the system.

Similar to the frame theoretic approach, the underlying assumption in case theory is that words evoke certain contextual expectations to the hearer, based on his personal experiences and his knowledge on stereotypic situations. In our system, this assumption is adopted in that we use case descriptions not only for verifying syntactic hypotheses, but also for syntactic and semantic predictions about the rest of the sentence. This top-down aspect plays an essential role not only in the semantic component but in the whole recognition process.

3. Semantic reasoning in EVAR

In our speech understanding system, the semantic analysis as defined above comprises the following tasks:
- resolution of lexical ambiguities
- interpretation of constituents with respect to their semantic features
- choice between alternative syntactic hypotheses and between alternative interpretations of constituents
- revelation of semantic anomalies due to recognition errors
- representation of the case structure
- inference of expectations on the rest of the sentence.

These problems are solved by three fundamental operations of the semantics module: *local interpretation* by unification of semantic features, *contextual interpretation* by case frame analysis, and *top-down hypotheses*.

3.1 Local interpretation of constituents

One of the main tasks of the module consists in mapping syntactic structures (hypotheses) to caseframe instances. As this mapping essentially relies on semantic features, the features of a phrase have to be determined first. On the one hand, this means resolution of lexical ambiguities, on the other hand, this process supports the choice between alternative word and structural hypotheses. The principle is to reduce lexical ambiguities by selectional features of the phrase heads that constrain dependent words and phrases. To determine the features of a phrase, all meaning alternatives of its constituents are unified and tested for compatibility. The test yields a rating that is the higher, the more constituents are compatible with the nucleus class. Of all possible feature combinations, the one with the highest consistency is chosen. The semantic consistency rating of a group can also be regarded as a measure for the plausibility of a syntactic hypothesis. As low semantic ratings may result from grouping wrong word hypotheses, a search for alternative word and constituent hypotheses may be reasonable in an area with bad semantic consistency.

The combinatoric constraints of words are expressed in the dictionary by the feature SELECTION. The system of semantic classes (features) is organized in a conceptual hierarchy, thus, with a given class selected by the phrase head all its subclasses are accepted as compatible. The system presently used consists of about 110 semantic features and is represented as a concept hierarchy in the network formalism.

3.2 Contextual interpretation

When constituents are locally interpreted, they are matched to the caseframes of some verbal groups in order to decide which constituents fit together and to represent their functional relationships. Usually there are different verb frames for a verb corresponding to its alternative meanings. The assumption is that the frame for the intended meaning will be the one that can fill most of its case slots.

The mapping of a semantically interpreted phrase structure to a caseframe is accomplished by three different matching functions. The syntax module produces syntactic structure hypotheses that are represented as network instances. Due to competing and erroneous word hypotheses and structural ambiguities there will be competing syntactic structures as well. Every syntactic hypothesis has a score to reflect its reliability and importance. Depending on whether a complete and spanning sentence hypothesis could be found, one of two matching functions is selected: *Frame Sentence Match* takes a good scoring sentence hypothesis, the immediate constituents of which have already been interpreted, and tries to match them to cases in the alternative frames of the head verb. Matching criteria are the constituent type that is required for a certain case and the selectional restrictions imposed by the verb.

The second version *(Frame Constituents Match)* has been implemented in order to cope with only partially recognized sentence structures, ie. with isolated constituents. It is expected that complete (and completely recognized) sentences more likely tend to be the exception in spoken dialogue, and that it is advantageous to envolve semantic interpretation as soon as possible. In this case, the frames of the best scoring verbal groups are matched to the best scoring constituent hypotheses.

For every successful configuration of a frame and filling constituents a frame instance is constructed with case attributes filled by the fitting constituents.

The matching process yields plausibility scores for the embedding of constituents into all alternative caseframes that may represent different meanings of the (assumed) head verb. The score is a function of different factors: the number of obligatory slots that could be filled, reliability scores from the other modules, consistency ratings of the constituents, fulfilment of selectional restrictions, the relative length of the time intervall (in the speech signal) not covered by the hypothesis.

The valency structure providing only a minimum framework for a sentence, a third interpretation function is needed to evaluate the functional relations of additional modifiers not constrained by valency. It mainly rests on the semantic properties of the 'functional words', that is prepositions and conjunctions, and of adverbs. Their semantic classes (eg. CAUSE, DIRECTION, SINCE) characterize the relation of prepositional and adverbial groups and subordinate clauses to the main clause.

3.3 Top-down analysis

Motivation The analysis so far can only be successful if a verb was uttered by the user that was also recognized with a satisfying certainty by the word recognition module. This is a very hard restriction for the user (to avoid for example elliptical constructions without an explicit articulation of a just mentioned verb) as also for the word recognition of the system.

The special problem with spoken natural language is that you will never have the really uttered string of word hypotheses which covers the whole speech signal and is furthermore syntactic correct. On the other hand it is likely that with all the generated word hypotheses there would be many possibilities of chaining some of them to such a string. So the system will neither find out if a word was uttered that isn't known to it nor that an ellipsis was uttered. That could be found only in written language, for example by communicating with the user by a terminal. But analyzing spoken utterances in a dialogue there would always be wrong alternatives to the unknown or missing word or the missing syntactic constituent.

This fact implies that it isn't possible to restrict the user to a certain range of speech, for example to formulate only complete sentences containing at least a subject and a verb. Whether any of such given restricting rules are violated is almost impossible to discover.

Besides this 'technical' point of view our system should 'behave' like a normal human communication partner, ie. it should be able to handle all formulations that are normally used in an information dialogue between two human partners.

Example:

U1: When does the next train leave for Hamburg?
S1: (there leaves one) At 12:15 hours.
U2: And (is there another one) a little bit later?
S2: That is the last (train to Hamburg) for today.

Such elliptical sentence structures (in which not only the verb is possibly missing but also a noun group such as in S2) prevent unnecessary redundancy and effect the conversation becoming more natural and fluent.

Top-down Hypotheses of Verbs In addition to the former described Frame Constituent Match, a kind of bottom-up analysis, a method is developed to analyze a spoken utterance without beginning with the verb of the sentence. Also this method is based on the valency theory (see above). Here we try to conclude from a set of constituent hypotheses produced by the syntax module to a set of possible verbframes containing slots for some of the found constituents which should not be competing with regard to the speech signal.

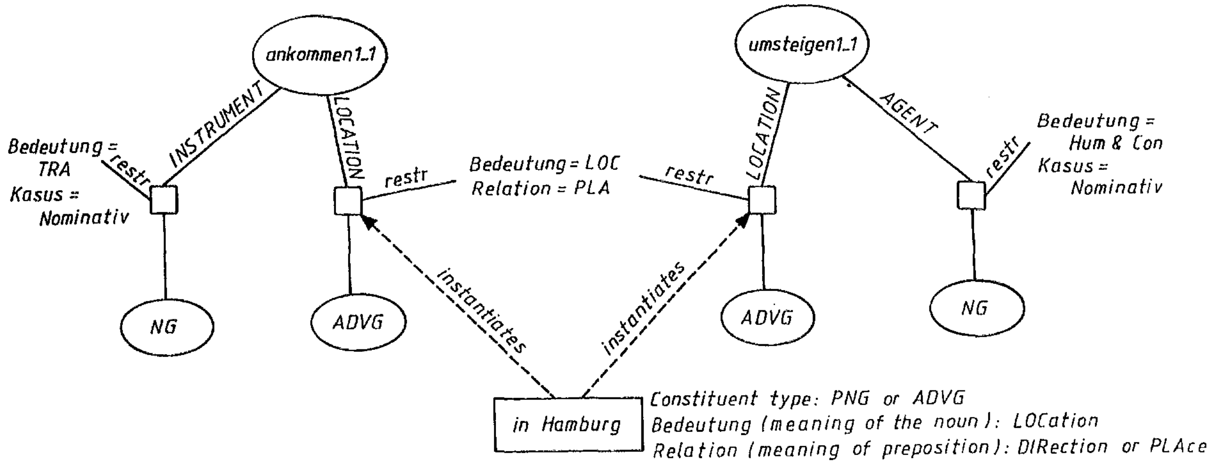
Therefore it was necessary to organize the database containing the verbframes in a way that the actants (represented as attributes of the concept verb in a semantic network) of the verb (the concept) could be attained not only by seeking the verb and its information, but also in a direct way without knowing the affiliated concept.

In German constituents have four selective features that can be used to restrict the number of the possible candidates for an attribute:

- the type of the constituent (for example noun group or prepositional group)
- semantic class which the constituent can be an instance of
- if the constituent is a prepositional or adverbial group the preposition respectively the semantic class of the preposition of the group
- the case of the noun of the constituent (if any noun is present).

For generating top-down hypotheses of verbs the last feature will not be used, because in German the endings which determine the case of a noun are all similar and so are the inflected word-forms of one lexeme. It is supposed (and partly shown by experiments) that the recognition and distinction of such word-forms is not reliable enough to base the further analysis on it. It would better serve for the verification of so far found syntactic and semantic hypotheses.

Example:



"ankommen1_1" corresponds to "arrive" in the meaning of "The train arrives at Hamburg." "umsteigen1_1" corresponds to "change" in the meaning of "I changed the train in Hamburg." The prepositional group (PNG) "in Hamburg" can be interpreted as the LOCATION attribute of "ankommen1_1" or of "umsteigen1_1".

Another problem with the lexicon is that it mustn't contain lexemes for many applications in order to reduce the possibilities of 'correct' verbframes. Although the semantics module in EVAR should be independent of a specific task domain it is not realistic to permit always all meanings of the whole lexicon for the semantic analysis. Therefore it is intended to use for the first step of analysis only a part of the lexicon which is locally determined by the pragmatic module and the dialogue module, dependent on the dialogue context and the expectations for the next dialogue step. Both modules together have the 'knowledge' about the world, as far as it is needed, the specific domain and the linguistic and situative context of the dialogue.

For the so far accomplished experiments two different verb lexicons were used. They were generated in a heuristic way limitating the whole range of our domain independent lexicon to a more or less restricted task domain. This was done prior to the analysis because up to now the pragmatic module is not realized. One of these lexicons contains only verbs that are used in our application 'Intercity Train Information'.

Other Top-down Hypotheses There are other possibilities too to generate top-down hypotheses in the semantics module:

- We try to reduce the number of the word hypotheses by first seeking semantically compatible word groups (they need not to be adjacent, but must not be competing). With this method the head verb and also descriptions for the syntactic realization of its attributes can be predicted.
- Another type of top-down hypotheses could be generated by seeking missing ie. not yet instantiated attributes of a verbframe, eg. "The train leaves *for Hamburg*."
- Sometimes the meaning of a sentence does not bear on the head verb but on a noun in that sentence, for example "Is there a good *connection* from Munich to Hamburg tomorrow morning." In such cases it regards a nounframe instead of a verbframe assuming that the head verb is performative like "ask", "excuse" and "must" or could be combined with nearly every noun like "have", "be" and "become".
- There is always the possibility to limitate the range of the speech signal for the top-down hypotheses: They only have to be sought where the so far found hypotheses are not. In addition information about word order in German sentences could often be used to restrict the possible range for a certain sentence part further.

4. Outlook

Experiments with the so far implemented semantics module indicate that without considering the dialogue context the semantic analysis will produce too many hypotheses. Therefore it will be necessary to take account of it with the further developments by making pragmatic predictions about the following user utterances.

With 'knowledge of the world', a special user model which describes all assumptions about the user and his intentions, and a memory about the course of the dialogue it is possible to predict the semantic and syntactic structure of the next user utterance, and also the words which can appear in this structure.

References

/1/: R.J.Brachman: A STRUCTURAL PARADIGM FOR REPRESENTING KNOWLEDGE. BBN Rep. No 3605. Revised version of Ph.D. Thesis, Harvard University.1978.

/2/: A.Brietzmann: SEMANTISCHE UND PRAGMATISCHE ANALYSE IM ERLANGER SPRACHERKENNUNGSPROJEKT. Dissertation. Arbeitsberichte des Instituts fuer Mathematische Maschinen und Datenverarbeitung (IMMD), Band 17(5), Erlangen.1984.

/3/: C.J.Fillmore: The grammar of hitting and breaking. WORKING PAPERS IN LINGUISTICS 1, The Ohio State University RF Project 2218-c, Report 1. In READINGS IN ENGLISH TRANSFORMATIONAL GRAMMAR, R.A.Jacobs & P.S.Rosenbaum (eds.). Waltham, Mass.1967.

/4/: H.Niemann, A.Brietzmann, R.Muehlfeld, P.Regel, E.G.Schukat: The Speech Understanding and Dialog System EVAR. In: NEW SYSTEMS AND ARCHITECTURES FOR AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS, R.de Mori & C.Y.Suen (eds.). NATO ASI Series F16. Berlin etc: Springer.271-302.1985.

/5/: L.Tesniere: ELEMENTS DE SYNTAXE STRUCTURALE. 2nd edition. Paris.1966.

This research was supported by the German Ministry of Research and Technology BMFT (in part by the joint project speech understanding in cooperation with Siemens AG, Muenchen).