An Application of Bayesian Networks to Lexical Ambiguity Resolution

Eduardo de Paiva Alves Furugoori Teiji
Department of Computer Science University of Electro-Communications

2 G－6

1 Introduction

The system described in the present thesis is able to syntacticly analize and identify the semantic cases and correct meaning for each word of the input sentence. The syntactic analysis is based on a formal grammar and is designed to identify the elements of the sentence. The semantic analysis is made using information about possible meanings and classification in a taxonomic hierarchy and semantic preferences. The two processes are interconnected and work in parallel since syntactic information may help semantic analysis and the other way round.

The model in which the system described in the present work was based is a Bayesian (Belief) Network (Pearl,1988) for the net structure makes possible both rule based reasoning and context information handling, while the probabilistic nature is adequate for treating conflicting and incomplete information and exceptions as well.

2 System Description

In this section, we describe the system conceptually. From the analysis of the problem, a network is designed, consisting of a syntactic and a semantic sub-networks. The way the different parts of the network are connected as well as how we chose to input the sentence are also presented.

2.1 The Syntactic Sub-Network

The input to the system consists of the identification of the position in which the words included in the lexicon occur as well as their possible grammatical categories. In the sentence:

The wind moves the boat

the input consist of the nodes: *the*1 *wind*2 *moves*3 *the*4 *boat*5 *art*1 *noun*2 *verb*2 *verb*3 *noun*3 *art*4 *noun*5 which stand for the association of words with the positions and categories.

The **Syntactic Sub-network** was specified from a context free grammar. In this sub-network, the different possible grammatical categories for each position of the input are grouped according to the rules of the grammar to form diagnostic nodes which correspond to the possible parsing of the input sentence. The links here may be seem as application of the productions of the grammar. This sub-network is connected to the semantic sub-network as it's further explained, making possible the exchange of information between the two sub-networks. In (Eizirik,1993), an algorithm for automatic generation of the network from a grammar can be found.

From the possible parsings of a input sentence, the noun and verb phrases are identified. Notice that some information on the possible associations are available from the input, such as *windNP*1 which is possible if the input has *wind*1 for instance. The syntactic parser interacts with the semantic through these nodes, so that some will be ruled out during the process.

Figure 1 depicts an extract from the **Syntactic Sub-network** relevant to the interpretation of the sentence *The wind moves the boat.* It shows how the grammatical categories are inherited from the input and the successively grouped to form an adequate syntactic diagnostic to the sentence. Notice that the nodes which relate the words from the input to the noun phrases are are directly connected to those allowing interaction between the two processes during the simulation.

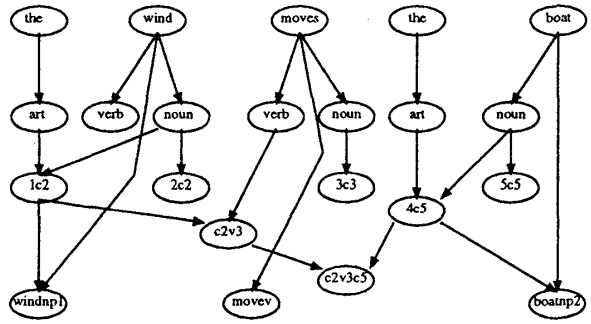

Fig. 1: Extract from the **Syntactic Sub-network**

2.2 The Semantic Sub-network

The **Semantic sub-network** classifies the noun phrases and verb-phrases and from this classification selects the possible semantic structures that combine these phrases into a sentence. It may be analyzed into two parts, the **taxonomic hierarchy** and the **semantic parser.**

2.2.1 The Taxonomic Hierarchy

The **Taxonomic Hierarchy** is a network in which the noun phrases are associated with categories which are hierarchically organized according to generality. For instance, *life* subsumes *animal* which subsumes *human.* The categories are those which allow the identification of the noun phrases with the fillers of the cases specified in the semantic parser.

The orientation of the links from the more specific to the more general characterizes inheritance, since if

the more general category is a filler for a verb case, the more specific one is also a filler for this case.

2.2.2 Semantic Parser

The second part of the semantic sub-network is the **Semantic Parser**. It identifies the existent dependencies in sentences using *Conceptual Structures* (Schank, 1977).

The diagnostic nodes store the information of how the concepts are related, that is, what role the concepts represented by the noun phases fill and which primitive the verb expresses. The links indicate which category is required for a given action.

For instance, in the case of semantic diagnostics, the occurrence of a node *ag atrans obj* depends on the occurrence of *human* on the first noun phrase, of either *give* or *lose* as the verb phrase and *thing* in the second noun phrase.

The cases correspond to the *Conceptual Cases: agent, directive/recipient* and *instrument.* The verbs represented in the network were chosen to be atomic, that is, correspond *Acts* with a single*primitive* and *conceptual case* were only represented when explicitly mentioned.

3 Results and Analysis

In the present section, the results of the simulations are presented, and the perform of the system in ambiguity resolution is analyzed.

The tests concentrated on the **Semantic Sub-Network** because tests on the **Syntactic Sub-Network** can be found in (Eizirik, 1993) which show the model's adequacy to rule application handling. The tests were made attributing a fixed value to the input nodes and then using the stochastic simmulation to determine the most probable configuration to the network, given the conditional probabilities. To speed up the process, simulated annealing was introduced, so that the system starts from a neutral configuration, and converges to the most probable configuration.

Besides the activation of the adequate semantic diagnostic, determining both the correct sense of the verb as well as the roles played by each noun phrase, notice also the activation of the categories corresponding to the noun phrases, as can be seen in the following example. In the picture, only the units relevant to the simulation of the sentence *John feels the heat* are represented.

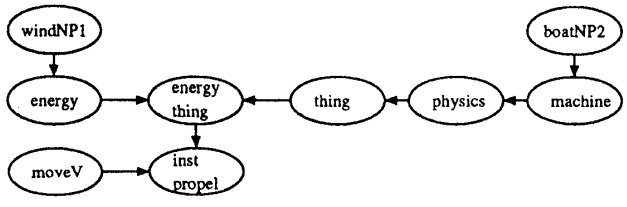

Fig.2: Extract from the **Semantic Sub-Network**

The identification of terms with categories disposed in a hierarchy was used to relate the nuom phases with the value restriction necessary to the interpretation of the sentence. The inheritance mechanism allowed the storage of semantic information in a non-linear hierarchy.

The cases of multiple inheritance, which happen in cases of structural ambiguity or context, may be handled with adequate attribution of probabilities. The probabilistic model also was adequate for handling exceptions. The representation using identical links was adequate since the localization in the net gave the extra information necessary to the interpretation.

References

[1] Eizirik, L. M. R., Barbosa and V. C., Mendes, S. B. T. (1993). Bayesian Network Approach to Lexical Disambiguition Problem, *Congitive Science,* 17(2), Ablex Publishing Corporation, Norwood, NJ.

[2] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufman Pub. Inc, San Mateo, CA.

[3] Schank,R.C. and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding* Lawrence Erlbaum Associates, Hillsdale, NJ.