A SEMANTIC ANALYSER OF NATURAL ITALIAN SENTENCES

M. Del Canto, F. Fusconi, L. Stringa

Elettronica San Giorgio - ELSAG S.p.A.
Via Hermada, 6 - 16154 Genova - Italy

This paper presents the analyser for type-written Italian sentences used in the LISA system.

This system has been developed in ELSAG within the researches on Natural Language Processing aimed at making the dialogue with the user easier and more graceful.

The analyser was designed to accept input sentences without any constraint on how they are formed (e.g. in active or passive form, with a variable number and position of complements, etc.), and to accept such irregularities as ellipsis, idioms and small grammatical errors. The present version works for simple sentences that are introduced one at a time.

The output of the system is an internal conceptual representation that we defined, according to Schank, "conceptualization". It represents the meaning of the sentence in a non-ambiguous way and we can define it as an organized set of predicate-argument lists, each of which can have some modifiers. One feature of the conceptualization format is its generality, which permits to easily adapt the system to any new application.

The analyser makes use of a dictionary that includes a lexicon, a set of syntactic descriptions and a set of semantic descriptions.

The lexicon, in the version presently implemented, contains about 2500 words that permit the user to introduce sentences (statements or questions or answers) con erning people characteristics like age, profession, health and their actions and relations with the external world, like to go, to come, to travel, to give, to take (TRANSFER), to know (KNOWLEDGE), to speak, to say (COMMUNICATION), to own, to contain (RELATION), etc. The lexicon is broken down into several sections in order to optimize the memory usage according to an exhaustive study of a root-ending representation of italian words.

The syntactic descriptions are directly related to lexical entries and contain attributes like gender and number of the nouns or like mood and tense of the verbs.

The semantic descriptions represent the conceptual entities related to lexical entries; they mainly contain informations about the consistency between predicates (named "operators") and their arguments and provide a classification of the conceptual entities.

The analysis process works in three main steps: the lexical analysis, a bottom-up recognition of "syntactic and semantic Groups" and a top-down insertion of these Groups in the semantic structure (conceptualization).

The lexical analysis recognizes the single words in the input string and compares them with the lexicon components in order to recover all possible interpretations. The sequence of the syntactic and semantic descriptions obtained forms the output of this step. Unknown terms are neglected at this point, but the dialogue controlling module is informed about that.

All the characteristics derived from the input are analysed in the next-step and organized in Groups based on the most meaningful terms (Verbs, Nouns, Adverbs). Each particular type of construction, corresponding to a Group, is recognized by an independent "specialist subgrammar", which selects only the relevant portion of the input.

The last step of the analysis relates the elements in the previously built Groups with the roles in the conceptual structure (conceptualization).

This structure in its main part ("nucleus") is determined by the verbs in the Verb-Group; some other roles, suggested by the "expectations" in the Noun-Groups and in the Adverb-Groups can be associated to the nucleus through RELATION operators. The main criterion that guides the association between Groups and conceptualization roles is based on expectations partly related to the verb and partly pre-established.

A semantic approach characterizes this method of analysis: in fact a syntactic representation of the input is not attempted and the semantic descriptions of concepts are directly accessible.

This makes it easier to guide the analysis towards the internal representation, at the same time reducing the number of the alternatives generated and consequently also the problem of dealing with ambiguities.

A further fundamental feature is the co-operation between bottom-up and top-down techniques in the organization of the input in Groups and in the filling of the output structure. The first of them is best suited to deal with grammatical deviations, incorrect inflexions and fragmentary utterances, also because of the manifold scanning and non rigorous constraints on the word positions.

At this level pattern-matching mechanisms can also be used to handle idioms and others fixed phrases; the recognition of these forms may need both syntactic and semantic descriptions.

The top-down approach in the last step is suitable to make the conceptualization a standard, manageable, structure, which could be easily adapted, if necessary, to specific needs.

This method of analysis proved to be a suitable one within a system capable of conducting a dialogue through non--constrained input sentences, related to a general and flexible knowledge representation.