INTEGRATING SEMANTICS AND FLEXIBLE SYNTAX BY EXPLOITING ISOMORPHISM BETWEEN GRAMMATICAL AND SEMANTICAL RELATIONS

Morena Danieli, Franco Ferrara, Roberto Gemello, Claudio Rullent

CSELT - Centro Studi e Laboratori Telecomunicazioni - Via G.Reiss Romoli 274, 10148 Torino, ITALY

. ABSTRACT

This work concerns integration between syntax and semantics. Syntactic and semantic activities rely on separate bodies of knowledges. Integration is obtained by exploiting the isomorphism between grammatical relations (among immediate constituents) and conceptual relations, thanks to a limited set of formal mapping rules. Syntactic analysis does not construct all the explicit parse trees but just a graph that represents all the plausible grammatical relations among immediate constituents. Such graph gives the semantic interpreter, based on Conceptual Graphs formalism, the discriminative power required to establish conceptual relations.

1. INTRODUCTION

In the field of automatic natural language understanding, the problem of connecting syntax and semantics has been faced in three different ways.

Some authors are persuaded that understanding natural language requires no use of syntactic knowledge. They claim that semantic representation can be built directly from the surface string, without the help of almost any syntactic source (1).

Other authors proposed highly syntactic systems, starting from the idea that the representation of the syntactic structure is preliminary to the understanding process (2).

While the work of this second group of researchers was concerned mainly with the understanding of individual sentences, the work of the partisans of semantics was about the understanding of whole texts.

This shifting of attention substained the idea that syntax and semantics should be used in an integrated way. Most researchers have thought that semantics and syntax should be integrated with respect to both the representation and the processing (3); others have claimed that it is more efficient to build a full-blooded syntactic representation during the parsing process (4).

Our approach shares some communalities with the last position. We reckon that semantic and syntactic processes should rely on separate knowledge bodies. Our effort is mainly focused on the realization of the integration by exploiting the isomorphism between syntactic structures and semantic representations, rather than by making syntactic and semantic processes interact, as it happens in previous integrated parsers (5). The idea of isomorphism is not carried out through one-to-one correspondence between syntactic rules and semantic ones - as in Montague-inspired parsers (6), but by mapping in a formal way grammatical and conceptual relations. The use of grammatical relations as intermediate level between syntax and semantics was also adopted in the KING KONG parser (7), but this system is still more near to the position which wants the representation of syntax and semantics as well as their processes to interact, while our choice is to maintain separate these different sources of knowledge.

The subsequent paragraphs describe how this hypothesis works in SHEILA (Syntax Helping Expectations In Language Analysis), a prototype developed at CSELT laboratories (Turin, Italy). The aim of SHEILA is to analyze and to extract relevant information from news (coming from the Italian news agency "ANSA"). The system is initially being applied to texts describing variations in the top-management of commercial societies; it has been fully implemented on a Symbolics Lisp machine. SHEILA takes advantage both from the use of expectations and from the combination of the results of a non-conventional syntactic analysis with the activity of a surface semantic analysis, based on the formalism of conceptual graphs (8). In this paper we describe just the principles which guide the integration between syntax and semantics. SHEILA correctly analyzes a set of thirty news, generating for each of them a set of records for a relational data base.

2. THE PROBLEM AND OUR PROPOSAL

In text understanding systems syntax and semantics have almost always been dealt with integration of their processing. Usually this kind of

(1) See the system IPP [Schank 80].
(2) The LUNAR system [Woods 72] is a classical example.
(3) An example is the Conceptual Analyzer [Birnbaum 81].
(4) See MOPTRANS [Lytinen 85].

(5) See PSLI3 [Frederking 85], FIDO [Lesmo 85] and WEDNESDAY-2 [Stock 86].
(6) See ABSITY [Hirst 84].
(7) See [BAYER 85]
(8) See [Sowa 84] and also the fourth paragraph below.

systems are semantic driven and they do only local syntactic checks during analysis. Doing local syntactic checks only involves little amount of syntactic knowledge and that is misleading in solving problems as anaphoric reference, prepositional attachment, conjunction and so on.

In a different approach the integration has been realized during the syntactic structure representation construction: the syntactic parser makes use of semantic information to handle structural ambiguities.

The questioning done by the syntactic analyzer to the semantic component aims to cut down the number of parse trees, but very many rules are required for this questioning, which has always been the most domain-dependent part of natural language understanding systems.

In designing SHEILA we chose another way of integrating syntax with semantics. The basic schema may look rather classic: the system produces a syntactic analysis of the text, driven on the basis of purely syntactic knowledge. The semantic analyzer checks the syntactic output to see if the semantic relations among words are supported by it.

But a classical syntax-first analysis is highly inefficient. It cannot solve structural ambiguities without the help of any semantic source and that leads to an explosion of the number of syntactic parse trees, some of them representing artificial syntactic ambiguities. So there are two problems: reducing the explosion of ambiguities and determining how semantic patterns for each word interact with syntax.

Our proposal faces these problems through the original combination of two key ideas, i.e.:
1) a flexible syntactic analysis, which is performed by constructing not all the explicit parse trees but just a graph, representing all the plausible grammatical relations among immediate constituents;
2) a formal way of interaction between syntax and semantics exploiting the isomorphism between syntactic structures (grammatical relations among immediate constituents) and semantic ones (conceptual relations).

Such flexible syntactic analysis gains a discriminative power (sufficient for aiding semantics in solving ambiguities) and avoids the explosion in the parse trees number. Furthermore, the mapping between grammatical and conceptual relations can be defined through a limited set of formal rules.

3. THE SYNTACTIC ANALYSIS

Our system has the goal of generating a semantic structure that has to be consistent with the syntactic form used to convey it in the text. The aim of syntactic analysis is to support semantics.

A first activity performed by the syntactic analyzer is the recognition of constituents of the

phrase structure of text. This is done by applying a set of rewriting phrase structure rules for Italian language. These rules utilize the output of a previous morphological analysis that assigns to words morphological and lexical features (gender, number, lexical category and so on).

In this analysis phase the application of the syntactic rules is limited to the recognition of the basic constituents of the phrase structure of the sentences. A basic constituent (BC, henceforth) is a NP, a PP or a VP described at a minimal level of complexity. At this level the grammar does not include rules of the form "S --> NP - VP" or "NP --> NP - PP", but it does include all the rules which describe the internal structures of BCs at the lowest level of recursion.

Every BC has a head and may have one (or more) modifier. The head of a BC is the characteristic word, the word without which a group of words would fail to be an instance of that particular BC. So the head of a NP is a noun, that of a PP is a preposition, that of a VP is a verb, etc.(9). The head of a BC carries on all the morphological, syntactical and lexical features of the BC itself (10). Let us consider the sentence

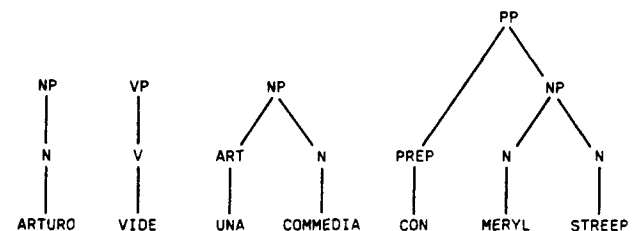(1) "Arturo vide una commedia con Meryl Streep.".

which may be interpreted both

(1.a) Arthur and Meryl Streep saw a play together

and

(1.b) Arthur saw Meryl Streep while she was working in a play.

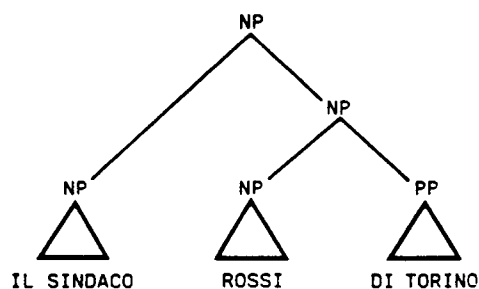At this first level of analysis (1) is rewritten as

---

(9) The case of PP constitutes a partial exception to this principle. In fact while for syntax is sufficient to know all the relevant information concerning the preposition, semantics also need to know the information concerning the head of the NP which forms the PP.

(10) This definition of head encompasses all constructions (endocentric and exocentric); it is closer to the traditional notion of governing categories than the definition given by Bloomfield [Bloomfield 35] in terms of distribution. See [Miller 85].

The output of this first step of syntactic analysis is a structure that includes the syntactic ambiguities which will be properly treated at the second level of analysis (11).

The second level of syntactic analysis has the goal of solving the problems about prepositional phrase attachment, noun phrase modification and conjunction and that of establishing grammatical relations among BCs (12). In the usual syntactic approach this activity, performed among more complex constituents, leads to the explosion of structural ambiguities. In our case the problem of handling ambiguity strongly arises: in fact the syntactic analyzer has been designed in order to treat a large variety of real texts which contain words out of their preferred grammatical order or which present elliptical constructions or, finally, which present very complex grammatical constructs. To reach such an adequacy we relax the grammar constraints, but that may cause the generation of artificial structural ambiguities (13). In order to solve this problem, we see all the groups of BCs having the same head as belonging to an equivalence class of constituents. Let us consider an example concerning this important point. In Italian the phrase "Il sindaco Rossi di Torino" ("The major Rossi of Turin") may involve some structural ambiguity if it has to be parsed without the help of semantic hints. In fact, this noun phrase can mean both that Rossi is the major of Turin and that Rossi is a major who comes from Turin. Performing a classical analysis this ambiguity generates two different structural descriptions. The first interpretation can be described as:

NP
  NP
    NP   NP   PP
  IL SINDACO ROSSI   DI TORINO

---

(11) At this level we have not so many ambiguities because the linguis-tic phenomena which cause them are still not faced. In this phase of analysis lexical ambiguity (involving uncertainty about the lexical category of a given word) only arises; this kind of ambiguity is treated by taking into account the syntagmatic relationships of the words in question; the analyzer keeps different interpretations for the ambiguity which can not be solved without semantics.

(12) Grammatical relations are primitive notions such as subject, object, complement and so on.

(13) The constraining power is provided setting up a structural homology between syntactic and semantic levels and performing the formal mapping between grammatical relations and conceptual relations.

while the second interpretation can be described as:

NP
  NP   NP
  NP   NP   PP
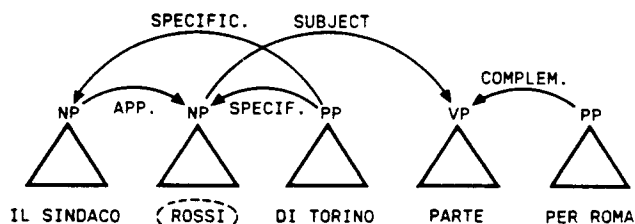IL SINDACO ROSSI DI TORINO

In our analysis we handle this problem starting from the consideration that in both the interpretations the NP "Rossi" is the head of the resulting structural unit. So the analyzer generates only one representation for the new construction in this way:

SPECIFICATION
  NP   APPOSITION   NP   SPECIFIC.   PP
IL SINDACO   (ROSSI)   DI TORINO
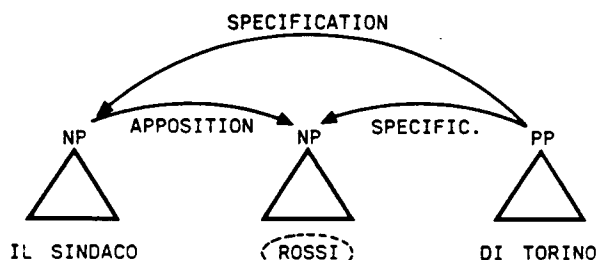
Now, let us consider this construction as being part of a sentence:

(2) "Il sindaco Rossi di Torino parte per Roma."
    "The major Rossi of Turin is leaving for Rome."

The ascription of grammatical relations among the phrases of this sentence requires the recognition of the NP "Il sindaco Rossi di Torino" as subject of the sentence and the PP "per Roma" as modifier of the VP. The detection of the subject relation does not necessarily involve the problem of structural ambiguity because this is limited at the relations between the two NPs and the first PP. So the analyzer gives the following description of the sentence:

SPECIFIC.   SUBJECT
  NP APP. NP SPECIF. PP   VP COMPLEM. PP
IL SINDACO (ROSSI) DI TORINO   PARTE   PER ROMA

Thanks to this treatment of ambiguity, the syntactic structure of this sentence can be described by only one representation, while a classical syntactic analysis would generate at least two repre-

280

sentations. Our single representation consists of a graph of BCs connected by grammatical relations, which are established unless syntactic knowledge guarantees that no constituent in the two classes can be connected by such relations. In this way the processing is efficient almost as in the case of complete parallelism between syntax and semantics and, in addition, there is complete compatibility with a parallel implementation.

Note that none of the possible interpretations has been lost: all them are passed to the semantic interpreter which operates the resolution of ambiguity taking into account both the connections between the BCs pointed out by syntactic analysis and the semantic plausibility of the proposed connections.

The resulting discriminative power of syntax is still sufficient for helping semantics in establishing the correct semantic relations among concepts denoted by words.

4. THE SEMANTIC ANALYSIS

Our working hypothesis is that we can represent the meaning of a text starting from the meanings of words and from the syntactic structure of the text.

We represent the surface semantic structure by conceptual graphs (14). A conceptual graph is an oriented bipartite graph with two kinds of nodes: concept nodes (representing entities) and conceptual relation nodes (representing semantic relations among concepts). A Type Hierarchy is defined over concepts.

The semantic information is distributed on words by means of canonical graphs, which describe concepts connoted by the words of the domain in terms of their semantic context; they represent the implicit pattern of relationships necessary for a semantically well-formed text. In each canonical graph we can distinguish a head (the main concept node of the canonical graph itself) and a semantic context (see figure 1). The Type Hierarchy is a taxonomy of domain concepts used to inherit semantic contexts and guide graph joins.

The aim of surface semantic analysis is to establish semantic relations among the head nodes of canonical graphs connoted by the words of text. First, the canonical graphs are activated (copied in the working memory); then the activated graphs are joined, superimposing context nodes on head nodes according with the Type Hierarchy; so relations are established among head concepts.

When establishing a semantic relation, the mapping with syntax allows the evaluation of its syntactic soundness: the syntactic analysis output

(14) The theory of Conceptual Graphs is presented by [Sowa, 1984]. This formalism is a generalization of various previous approaches to the representation of the semantic relations holding among words such as frames, semantic networks and conceptual dependency.

is checked to see if a grammatical relation supports the proposed semantic one. Otherwise the semantic relation is not established.

5. INTEGRATING SYNTAX AND SEMANTICS

During semantic analysis relations between concept nodes are established only if they are supported by the result of syntactic analysis.

Given a semantic relation, it is necessary to see if there is a corresponding grammatical relation. The correspondence between grammatical relations and semantic relations (mapping) is solved through the notion of head which has been introduced both in syntax (heads of BCs) and in semantics (heads of canonical graphs).

The semantic relations and the grammatical relations must relate to the same couple of lexical items; in other words such lexical items must be both the heads of the BCs (involved by the grammatical relation) and the heads of the conceptual graphs (involved by the semantic relation).

A semantic relation SR between two head nodes HNi and HNj, having as heads the words Wi and Wj, can only be established if:
1) there is a grammatical relation GR between two BCs, BCi and BCj, whose heads are Wi and Wj respectively.
2) semantic relation SR is compatible with the grammatical relation GR and with the set of features Fi and Fj associated to BCi and BCj.

Conditions are verified through the application of a mapping rule among a limited set. Each semantic relation inside a semantic context of a canonical conceptual graph is augmented with the indication of a mapping rule.

A mapping rule is a list of plausible grammatical relations that can correspond to the semantic relation.

In a mapping rule each grammatical relation can be constrained by an activation condition that relates to the morphologic and syntactic features of the involved BC classes.

5.1 An example

· Let us consider the example of the figure 2.

The join J1 of the head conceptual node HN1 with the context node CN2,1 of the head node HN2 causes a conceptual relation AGENT to be established between concept nodes HN1 and HN2. Such head concept nodes correspond to words W1 ("John") and W2 ("eats") at the lexical level.

Such conceptual relation has an associated mapping rule which requires a grammatical relation of a certain kind (e.g. "subject"). Such grammatical relation must have been established by syntactic analysis between two BCs having W1 and W2 as their heads. As that is the case of figure 2, the join J1 can be made.

Differently, join J4 between HN3 and CN2,1 can not be established as it would cause an AGENT relation between conceptual nodes HN2 ("eat") and HN3 ("chicken"); such semantic relation is not supported by a suitable grammatical relation. In fact there is a grammatical relation between BC2 and BC3, but it is not the correct one because the grammatical relation "object" can not correspond to the semantic relation AGENT.

To give an idea of the mapping rules, the MR-AGENT mapping rule is sketched. It is used to map the conceptual relation AGENT on the grammatical relation "subject" if the analyzed sentence is active or on the grammatical relation "agentive" if the sentence is passive:

MR-AGENT : subject   if   BC1 is ACTIVE and
                            BC1 and BC2 agree.

          agentive   if   BC1 is PASSIVE and
                            BC2 is a "by-phrase"

6. CONCLUSION

The SHEILA system has been presented as an attempt to solve the problem of integrating syntax and semantics. The authors propose that syntactic and semantic processes should rely on distinct bodies of knowledge and that the interaction between syntax and semantics should be obtained by exploiting, in a formal way, the isomorphism between syntactic and semantic structures. In order to avoid the lack of efficiency characterizing a syntax-first parser, the authors have designed a flexible syntax which, without exploding the structural ambiguities, supplies semantic interpreter with knowledge about syntactic connections between the words occurring in the text. The isomorphism between syntax and semantics is accounted into a limited set of formal mapping rules and conditions. Prepositional phrase attachment, apposition, determination of conjunction's scope and modification of a NP through other NPs are dealt in a satisfactory way both from a syntactical and from a semantical point of view. Other complex linguistic phenomena (as anaphora, quantification and ellipsis) requires a more extensive use of heuristics. The future work will concentrate on these specific aspects in order to check the adequacy of the hypothesis of isomorphism between syntactic and semantic structures to larger fragments of the Italian language.

REFERENCES

[Bayer 85] Bayer, S., Joseph, L., Kalish, C., Grammatical Relations as The Basis for NL Parsing and Text Understanding. **Proc. 9th IJCAI**, Los Angeles, 1985, pp. 798-790.

[Birnbaum 81] Birnbaum, L. and Selfridge, M., Conceptual Analysis for Language. in Schank, R.C. and Riesbeck, C.K., (eds), **Inside Computers Understanding.** Lawrence Erlbaum Ass., 1981.

[Bloomfield 35] Bloomfield, Leonard, **Language,** Allen & Unwin, London 1935.

[Frederking 85] Frederking, R.E., Syntax and Semantics in NL Parsers. Technical Report 133, Carnegie-Mellon, Dept. of Computer Science, May 1985.

[Hirst 84] Hirst, G.J., **Semantic Interpretation Against Ambiguity.** Brown University, Ph.D., 1984

[Lesmo 85] Lesmo, L. and Torasso, P. Weighted, Interaction of Syntax and Semantics in NL Analysis. **Proc. 9th IJCAI**, Los Angeles, 1985, 772-778.

[Lytinen 85] Lytinen, S.L., Integrating Syntax and Semantics., **Proc. Theoretical and Methodological Issues in MT for NLs**, Hamilton, 1985, 167-178.

[Miller 85] Miller, J., **Semantics and Syntax,** Cambridge Univ. Press, Cambridge (U.K.), 1985.

[Stock 86] Stock, O., Dynamic Unification in Lexically Based Parsing., **Proc. 7th ECAI**, Brighton, 1986, 212-221.

[Sowa 84] Sowa, J.F., **Conceptual Structures.** Addison Wesley, 1984.

[Woods 72] Woods, W.A., An Experimental Parsing System for Transition Network Grammars. Technical Report 2362, Bolt Beranek and Newman Inc., 1972.

Fig.1 - The canonical graph of "eat" and that of "fork".

Fig.2 - Mapping aspects for the sentence "John eats a chicken with the fork". The syntactic level represents the graph of BCs that constitutes the two syntactic structures of the sentence. At the semantic level dotted arrows (⟺) stand for a join that is supported by syntax. The double arrows (⟸⟹) instead represents a join that is not supposed by syntax. In fact a mapping rule requires that the semantic relation "agent" must be supported by the grammatical relation "subject" (in an active sentence) and not by the "object" relation.