KNOWLEDGE EXTRACTION FROM TEXTS BY SINTESI

Fabio CIRAVEGNA
Paolo CAMPIA [1]
Alberto COLOGNESE [2]

CENTRO RICERCHE FIAT
strada Torino 50,
Orbassano (To),
Italy

ABSTRACT

In this paper we present SINTESI, a system for the knowledge extraction from Italian inputs, currently under development in our research centre. It is used on short descriptive diagnostic texts, in order to summarise their technical content and to build a knowledge base on faults. Often in these texts complex linguistic constructions like conjunctions, negations, ellipsis and anaphorae are involved. The presence of extragrammaticalities and of implicit knowledge is also frequent, especially because of the use of a sublanguage. SINTESI extracts the diagnostic information by performing a full text analysis; it is based on a semantics driven approach integrated by a general syntactic module and it is able to cope with the complexity of the (sub)language, maintaining both accuracy and robustness. Currently the system has been tested on about 1.000 texts and by a few users; in the near future it will be used by dozens of users every day.

1. INTRODUCTION

In the last years a great interest in the information retrieval from texts has grown; as a matter of fact the more the ability of memorising large quantities of data increases, the more the difficulties in extracting grows. The classical information retrieval approaches that consider the input only from a formal point of view are not powerful or user-friendly enough, as they are not able to cope with the real content of the texts; the quality of their results is generally poor. When a higher quality is needed, it is necessary to adopt methods derived from the field of the natural language processing, to extract some structured knowledge (as the objects mentioned and their relations). In this case different kinds of applications require different architectures. As a matter of fact, to extract information from news it is necessary to be able to cover a wide range of correct syntactic forms, but a little of extragrammaticalities. At the same time the general knowledge sources are more important than the domain dependent [Jacobs 88]. On the contrary, in the case of short technical texts (for example diagnostic reports) a wide syntactic coverage is not the main needing, but a large use of extragrammaticalities or a sublanguage must be taken into account [Liddy 91]. At the same time the domain dependent knowledge sources assume a main role, especially because of the presence of the implicit knowledge. In the latter case many of the approaches proposed in the field of NLP are not suitable or powerful enough, as they should guarantee three main features: efficiency, robustness and accuracy.

Efficiency is necessary because the input is generally long (more than one sentence), requiring a strong treatment of phenomena such as anaphora and ellipsis. Moreover, the efficiency is important when a system must operate in real time.

Robustness is needed because of the use of a sublanguage, the presence of implicit knowledge and/or of ill-formed sentences.

Accuracy is needed because the objects involved in a technical description are generally very complex from a linguistical point of view (for example a car part name may be composed by even ten words), and these descriptions are generally affected by the problems of the sublanguage and of the implicit knowledge. Accuracy allows to resolve those problems by using not only the knowledge of the world, but the linguistic information, too.

Accuracy and robustness are difficult to obtain at the same time; many classical approaches to natural language processing guarantee accuracy but fail in robustness; other methods are robust, but not accurate. The following techniques have been proposed in the last years to cope with the ill-formedness [Kirtner 91]:

1.The addition of some special rules to a formal grammar [Weischedel 83];

2.The introduction of some grammar-independent syntax-driven [Mellish 89] or semantics driven rules [Kirtner 91].

3.The treatment of the ill-formedness as a correct form of a sublanguage [Lehrberger 86].

4.The Use of semantics driven approaches as caseframe parsing [Carbonell 84].

In this paper we present our experience in building a system to extract knowledge from short technical diagnostic texts; we adopted a full-text semantics driven analysis integrated by the use of a general syntactic parser. The ill-formed input is treated without introducing special rules or

[1] Grantee ATA; current address corso Croce 7, Torino.
[2] Grantee ATA; current address via Prospero 47, Grugliasco (To).

sublanguage concepts; as we will see, we use the advantages of the case-frame approach to obtain robustness, while the syntactic knowledge is used when accuracy is needed.

2. OVERVIEW OF SINTESI

SINTESI (Sistema INtegrato per TESti in Italiano) [Ciravegna 89-92] is a prototype for the knowledge extraction from Italian inputs, performing a full-text analysis. It is used on short descriptive technical texts (four or five sentences in seven or eight lines) containing complex linguistic constructions like conjunctions, negations, ellipsis and anaphorae. The use of extragrammatical language and of implicit knowledge is also frequent. Typical of our domain (car fault reports) are the complex object descriptions (a full car part description may involve the use of even ten words). The goal of SINTESI is to extract the diagnostic information (main fault, chain of causes, chain of effects, car parts involved, etc.) from each text and to build its semantic representation. In the rest of the paper we will show how the different knowledge sources contribute to the text analysis and how they contribute to guarantee the robustness, the efficiency and the accuracy.

3 THE SENTENCE ANALYSIS

SINTESI integrates five knowledge sources: lexical, syntactic, pragmatic, general semantic and world knowledge. The input texts is first pre-processed by a morphological analyser with the help of a dictionary (currently containing about 4.000 entries). A lexical-semantic analysis is then applied to recognise some special patterns such as ranges of data or numbers, measurements, chemical formulas, etc, through a context free grammar. Some special preliminary information ("the semantic markers") is also put in the sentence to help semantics in the following steps.

The rest of the analysis is based on a semantics driven approach that integrates in a flexible way two modules dedicated to the linguistic, two to the semantic and one to the pragmatic analysis (fig.1). The semantic modules lead the analysis in order to perform two main steps: object recognition and object linking. This separation is introduced in order to be able to recognize fragments, when a complete analysis is impossible. The syntactic modules are used to build the linguistic structure of the objects and of the whole sentence. The pragmatics is mainly used to control the object linking. Additional modules provide the discourse analysis and the correlation of the knowledge extracted in the different sentences of the text. The pragmatics and the additional modules are not discussed in this paper.

3.1 OBJECT RECOGNITION

The object recognition step processes the text from left to right trying to identify the objects in each sentence. It is a bottom-up task that uses four kind of knowledge: the semantic markers, the general semantic knowledge, the world knowledge and the syntactic knowledge. The schema of this step is shown in fig.2. The presence of the objects is shown by the **semantic markers** put by the lexicon. The markers activate an expectation in the semantic module that is used to control the analysis and guarantee the robustness. The syntactic module is then activated to recognise the structural form of each object without trying to build a full sentence representation. The **syntactic knowledge** [Campia et al 90] is represented by an independent grammar. The syntactic structure of the linguistic expressions is represented by dependency trees. This kind of representation was chosen because it makes the syntax-semantics interaction easier, the interdependencies among words being shown in a very clear manner. The syntactic analysis is done by a set of production rules in which the conditions test the current tree status, whereas the actions modify it. A semantic test is performed immediately after each syntactic action in order to improve the efficiency and reduce the use of the backtracking [Ciravegna 91a]. Each semantic test activates the semantic module which uses two kinds of knowledge to answer to the test and to build the semantic representation of the current object: the general and the world knowledge.

The **general semantic knowledge** is based on caseframes and contains the basic information to answer to the question (general description of objects, information about roles and role-fillers, etc.). The caseframe model was derived from the Entity Oriented Parsing approach [Hayes 84]. The information contained in the caseframes produces a first hint on the identity of the object. Performing a semantic test at this level means filling a role in a caseframe. The precise identity of all the already known objects is contained in the **world knowledge** and it is formed by the syntactic and semantic descriptions of each object [Lesmo 90,91][3].Only the objects interesting for the knowledge extraction are contained in this structure. Performing a semantic test at this level means matching the syntactic representation on the contained descriptions via a set of structural rules. Every accepted connection between two words contributes in filling both the roles of the caseframe and to extract a semantic identity. The syntactic analysis of the current object description ends when the next word is not belonging to the current object description (for semantic, pragmatic or linguistic reasons). The object is then closed and the control is returned to the semantic

[3] The formalism was developed in collaboration with a group of the University of Torino.

controller that will try to identify another object, until the end of the sentence is found.

As example of the analysis of a very easy description of a car fault, consider: "Fissaggi della coppa [dell'] olio [del] motore con cricche" (literally: "Bolts of the slump [of the] oil [of the] engine with breaks"). The first semantic marker is put on "fissaggi" that activates an entity of type "car_part" (fig. 3a). At this step the role-expectation given by the caseframe shows that in the rest of the sentence we will probably have some other words specifying the current object and a fault associated to the car_part. The expectation given by the world knowledge is given by the following car part descriptions:

a. (fissaggio (perno (banco (motore))))
b. (fissaggio (coppa (olio (motore))))
c. (fissaggio (coppa (riciclaggio (olio))))
d. (fissaggio (ammortizzatori (anteriori (telaio))))

These descriptions are transformed in the corresponding dependency trees (see fig.3). The syntactic module is then activated and it tries to connect "coppa" to "fissaggio" through the preposition "del"; the connection is semantically acceptable because there is a caseframe plausibility, and the preposition is acceptable for the connection. Moreover there are domain based semantic descriptions that support the connection ('b' and 'c' only, because the others don't support the word "coppa"), so the semantic module is able to continue the current object building and the parser does the same. The analysis continues in the same way until the word "cricche" is found; "cricche" is pointing to another kind of entity (a fault), so the semantic controller stops the syntactic module. The fault description is then analysed in the same way. At the end of the object recognition we will have two objects: a car part (given by a caseframe and the 'b' description) and a fault.

SINTESI is able to cope with the loss of the determiners and prepositions in the description, because the dependency grammar shows only the structural relations between the different syntactic types; for example a typical rule shows that a determiner (or a noun) may be attached to a noun and not something like: "NP-> det+Noun". This kind of ill-formedness are then overcome by the formalism, without introducing metarules, sublanguage concepts or semantics driven rules.

At this level the interaction between syntax and semantics brings to the object recognition from a structural and semantic point of view when the object is present in the world knowledge base. When this identity is not known, it is possible to recognize the presence of the object by using only the syntactic module and the caseframe level of the knowledge base (role expectation). Even if it is not possible to derive a direct identity (for structural reasons), it is possible to recognize it by using the role-expectation coming from the other objects in the sentence. In this way, it is possible to maintain the accuracy on the identity and on the structure of the objects (if the description is already known), maintaining the robustness (when it is not known or it contains unknown words or unsolvable gaps).

Note that even if the object recognition is semantics driven, the approach is flexible: sometimes in fact it becomes syntax-driven; it happens to treat some special cases as the noun+ adjectives+ noun construction due to the loss of the preposition between the two nouns. These forms often give origin to nominal compounds, and, especially in conjunctions, bring to garden paths and different rules for the adjectives [Campia et al 90].

3.2 OBJECT LINKING

When all the objects of a sentence are identified by the object recognition step, a connection among them is tried, using the role-expectation contained in the caseframe level. There are two possible kinds of connection strategies: total linking and partial linking. The first one is tried when no failures were reported during the object recognition. It integrates Bottom-Up (BUS) and Top-Down (TDS) strategies in order to build the structure of the whole sentence. The **TDS** is an expectation-driven analysis of the connections among objects, driven by the main roles of some kinds of constituents (verbs, conjunctions, etc.). It is also driven by linguistic and pragmatic rules. The TDS is not executed in a left-to-right way and it is able to cope with some kinds of garden paths involved by conjunctions. The **BUS** is performed instead from left to right, and connects the objects not considered by the TDS. For every object a role-driven connection is tried with all the objects that are linguistically and semantically acceptable. A focus stack is used to control the connections. A score is given to every linking. This score is integrated by other evaluations coming from the domain specific knowledge. Different strategies on the connections are adopted according to different cases: it is possible that sometimes even some linguistically non acceptable relations may be accepted for pragmatical reasons.

When the object recognition analysis reported some problems, only the BUS is adopted in order to form some aggregates of objects. This aggregation is influenced by the content of the unknown or incorrect parts of the input. Connections among objects separated by obscure parts are considered unlikely. A classification of the unknown parts is done trying to apply some lexical, syntactic or semantic heuristic rules on the identity of the words contained (a main verb has a strong power of separation, a group of adjectives hasn't it, and so on).

In the example of section 3.1, the two objects (the car_part and the fault) are linked via the slot

"with_fault" of the first object. The connection is done by the BUS strategy because the lack of a verb in the sentence makes the system to hypothesize some gaps or ellipsis. The result of the analysis is shown in fig 3b.
The separation between object recognition and object linking guarantees the accuracy (when possible) and the robustness, by adopting different strategies according to the reliability rate of the object recognition. Note that even the TDS+BUS allows to cope with sentences that don't bring to a complete structure (i.e. it is not possible to connect some objects to any role in the sentence). It is possible because the BUS is always applied and it is this strategy that guarantees the robustness during the object linking.

4. ADVANTAGES OF THE SCHEMA

The illustrated strategy fits our requirements of accuracy, robustness and accuracy.
Accuracy is achieved because the syntax-semantics interaction during the object recognition and the TDS+BUS (during the object linking) bring to a full sentence structural and semantic definition when the input is correct.
Robustness is achieved during the object recognition because the syntactic analyser is able to cope with some kinds of ill-formed input (the loss of prepositions and determiners), without introducing extra-rules. In any case the analysis is still lead by semantics, so it is possible to force the syntactic module to accept an incorrect input. Moreover the semantic previsional ability makes us to cope with even unknown or incorrect object descriptions. It is possible by excluding the world knowledge and remaining at the caseframe level. The role-expectation driven analysis allows to detect the new descriptions and to propose their addition to the knowledge base.*
The robustness is also achieved during the object linking, because it is possible to adopt different strategies to cope with different rate of understanding or ill-formedness (unknown parts of the input, the lack of some fundamental constituents as the verbs, etc.).
Efficiency is guarantied because at the object recognition level each syntactic connection is immediately semantically tested, so the interaction is efficient. Moreover the tests at the world knowledge level are efficient, because they are reduced to a comparison between graphs (the current syntactic tree and the possible deep descriptions); it is particularly important in the reduction of the gaps involved by conjunctions [Ciravegna 91b].
In addition the separation between object recognition and object linking allows to insert between them an additional module called SKIMMER that improves the efficiency; it is a function that, given the interesting types of objects for the knowledge extraction (faults, car_parts, etc.) and the types of the object contained in the current sentence, decides if the sentence will bring new interesting information or not. If not the object linking and knowledge extraction are skipped. It is necessary to apply the skimmer after the end of the object linking (and not before) because a negation or a semantic modifier may change the meaning of an object. Moreover the anafora resolution is not affected by the skipping of a sentence, because the objects are already recognised.

5. CONCLUSIONS

This paper presents a system to extract knowledge from domain-oriented ill-formed Italian inputs. The purpose of the paper was to demonstrate how it is able to guarantee efficiency, robustness and accuracy. SINTESI is currently used to extract knowledge from technical diagnostic texts on car faults. From a linguistic point of view it is able to extract knowledge from sentences involving the use of noun phrases, verb phrases and prepositional phrases; the sentences may contain conjunctions, a limited set of garden paths and some kinds of subordinates. The system has two ways to operate: an on-line mode in which each new text is analysed in real time and the extracted knowledge is approved or refused by the user; an off-line mode to process the 40.000 texts that are already in a database. The extracted knowledge is used to generate search keys for the database, for statistical matters and to build a knowledge base on faults. One of the goal of the system is the transportability through the applications in the same domain. Currently SINTESI has been tested on about 1000 technical texts; the rate of the correctly extracted information was of about 85%. Many problems came from not currently supported forms, unknown objects or words, and complex garden paths. The system is able to process about 150 texts per hour running on a VAX 6510. It was developed by using the Nexpert Object tool and the C-language; it is now running in a DEC-VMS environment.
In the near future we will extend SINTESI in order to cover most of the linguistic forms that are still not covered. A method to extensively cope with the implicit knowledge is under development. Until now the system has been tested by few users, but it will be utilised by dozens of people with a rate of about 5.000 texts per year.

6. ACKNOWLEDGMENTS

We would like to thank dr. R. Tarditi (Fiat Research Centre), prof. L. Lesmo, dr. P. Terenziani and dr. V. Lombardo (University of Torino) for their help in the development of the

* It is then possible to build semiautomatically this part of the world knowledge base for the new applications.

world knowledge formalism and for their precious suggestions.

BIBLIOGRAPHY

[Campia et al 90]: Campia P., Colognese A.:"Organizzazione della conoscenza sintattica e interazione con la semantica in un sistema per la comprensione di testi"; Tesi di laurea, Torino 88

[Carbonell 84]: Carbonell J.G., Hayes P.J.:"Recovery strategies for Extragrammatical Language"; Computational Linguistics 10, 1984

[Ciravegna 89-92]: Ciravegna F.:"SINTESI"; four technical reports, Torino 1989, 90, 91, 92

[Ciravegna 91a]: Ciravegna F., Tarditi R., Campia P., Colognese A.: "Syntax and Semantics in a text interpretation system"; The RIAO91 Conference on intelligent text and image handling, Barcelona, April 1991

[Ciravegna 91b]: Ciravegna F., Campia P., Colognese A.:"The treatment of conjunctions in an information retrieval system", The Second Italian Conference on Artificial Intelligence, Palermo, October 1991

[Hayes 84]: Hayes P.J.:"Entity Oriented Parsing", COLING 84, 1984

[Jacobs 88]: Jacobs P., Rau L.: "Integrating Top-Down and Bottom-Up Strategies in a Text Processing System", 2nd Conference on Applied Natural Language Processing, Austin, 88

[Kirtner 91] Kirtner J.D., Lytinen S.L.: "ULINK: A Semantics-Driven Approach to Understanding Ungrammatical Input", AAAI-91, Anaheim 1991

[Lesmo 90,91]: Lesmo L., Terenziani P., Lombardo V.:"Relazione della convenzione di ricerca tra Universita' di Torino e Centro Ricerche Fiat", Torino 1990, 1991

[Lehrberger 86]: Lehrberger J.: "Sublanguage Analysis" in "Analyzing Language in Restricted Domains: Sublanguage Descriptions and Processing" edited by Grishman and Kittredge, LEA publ., London 1986

[Liddy 91]: Liddy E.D., Joergenson C.L., Sibert E., Yu E.S.:"Sublanguage grammar in natural language processing for an expert system"; the RIAO91 conference, Barcelona, 1991.

[Weischedel 83]: Weischedel R.M., Sondheimer N.D.:" Metarules as a Basis for Processing Ill-formed Input", in American Journal of Computational Linguistics, Vol 9, 1983

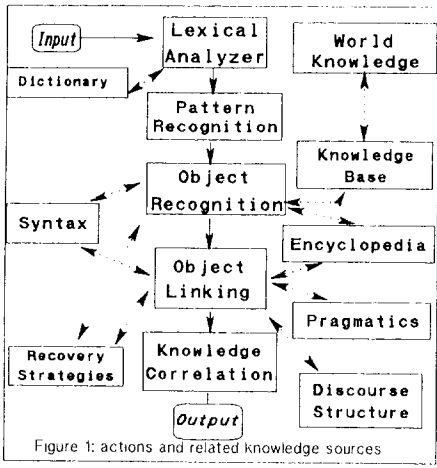

Figure 1: actions and related knowledge sources

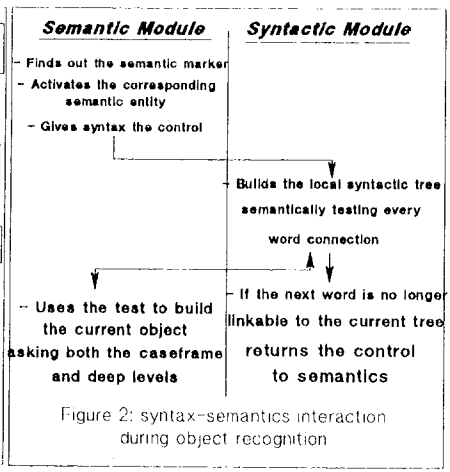

Figure 2: syntax-semantics interaction during object recognition

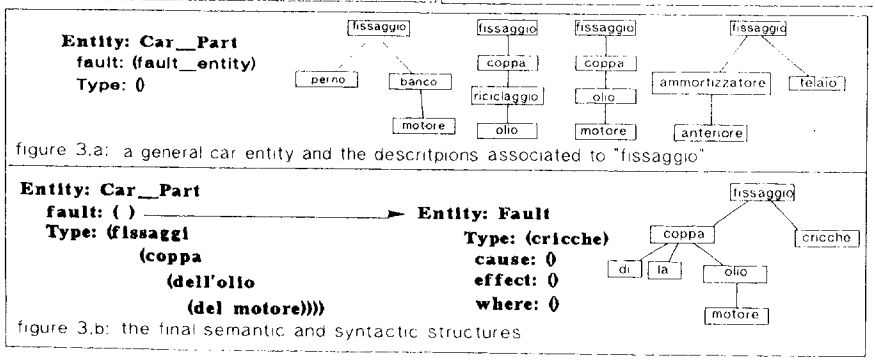

figure 3.a: a general car entity and the descritpions associated to "fissaggio"

figure 3.b: the final semantic and syntactic structures