Robust parsing of natural language descriptions expressed in telegraphic style

Michela Fasolo, Lorenzo Garbuio, Andrea Malanima
Centro Studi CERVEDomani, CERVED SpA, Corso Stati Uniti 14, 35020 Padova (Italy)

Nicola Guarino
Italian National Research Council, LADSEB-CNR Corso Stati Uniti 4, 35020 Padova (Italy)
email: guarino@ladseb.pd.cnr.it

1 Introduction

Sublanguages represent an important application area for NLU (Grishman and Kittredge,1986). Their syntactic simplicity and reduced semantic variability provide clear computational advantages. In the present paper we consider a sublanguage currently used for official publication of business activities which is characterized by a telegraphic style typical of commercial ads. Morphological and syntactic ill-formedness is very frequent within this sublanguage, hence a robust parser is a must.

The corpus we have considered was extracted from the on-line archives of the Italian Chambers of Commerce, and contains about 4 million descriptions of economic activities. They represent an important source of information about the structure of the Italian economy. Since our main goal is intelligent information retrieval, only a part of the information contained in the sentences is considered relevant. Basically, the kind of information we are interested in involves *nouns, prepositions* and *noun modifiers*, and involves verbs only in their nominalized or infinitive form.

The peculiarity of the parsing approach described in the paper consists in the fact that we limit the syntactic analysis to the elementary relationships occurring among these elements, discarding whatever is not recognized by the morphological analyzer and giving up the attempt to reconstruct the syntactic tree of the whole sentence.

2 The sublanguage

Examples of the sublanguage we are considering are reported below:

Produzione conto terzi olio di oliva.

*production (for) third parties (of) olive oil.[1]

Mulino. Produzione di foraggio cereali mais.

*Mill. Production of fodder cereals corn.

La produzione, l'importazione, l'esportazione, il commercio all'ingrosso e al dettaglio di prodotti tessili in fibre naturali e sintetiche e articoli per la casa in genere.

*Production, import, export, wholesale and retail trade of textile products of natural and synthetic fibres and articles for the home.

As we can see, the utterances are mostly formed by complex noun phrases rather than complete sentences, in which the number of constituents and the nesting degree may be rather high. Ellipses of prepositions and/or coordinating conjunctions are also frequent; the use of adjectives is relatively poor, and, when used, they generally bind the technical component of the information; the use of locutions with adjectival function is also common. Finally, the usage of verbs is quite scanty, and restricted to conversational parts; lexical mistakes are also recurrent.

In their basic version, these noun phrases are formed either by a single noun or by a group of the form N1-P-N2 (noun, preposition, noun) where N1 is associated with an economic activity, and N2 with the object of this activity linked by a preposition, which is often absent.

3 Syntactic analysis

The core of our approach is the extraction of *elementary syntactic relationships (ESR)* from a possibly ill-formed sentence. ESR's are described by a *definite clause grammar* (DCG), a fragment of which appears in Fig.1. Due to the presence of the special symbol *skip* in the right side of the rules, this grammar turns to be a special, very simple case of a *discontinuous grammar* (Dahl, 1989).

```
esr(na(N,A))          → na(N,A).          % noun-adjective
esr(npn(N1,P,N2))     → npn(N1, P,N2).    % noun-preposition-noun
esr(nn(N1,N2))        → nn(N1, N2).       % noun-noun
esr(ncn(N1,C,N2))     → ncn(N1,C,N2).     % noun-conjunction-noun

na(N,A)          → noun(N), adj(A) .
npn(N1, P, N2)   → noun(N1), skip, prep(P), possible_adjs, noun_group(N2).
nn(N1, N2)       → noun(N1), possible_adjs, noun(N2).
ncn(N1, C, N2)   → noun(N1), skip, conj(C), noun_group(N2).
noun_group(N)    → noun(N) | art(A), noun(N).
possible_adjs    → adjectives | [ ].
adjectives       → adj(A), possible_conj, adjectives.
possible_conj    → conj(C) | [ ].

Prolog expansion of the skip symbol:
skip(S0,S2):- append(S1,S2,S0).
```

Fig 1 A part of our DCG grammar

[1] The translation provided is rather literal, but reflects the ungrammatical, telegraphic style of the sublanguage.

The introduction of the *skip* allows us to ignore unknown or ill-formed words, and accounts in a very compact way for the positional variations of the structure elements.

4 The system

Input text first passes through a morphological analyzer similar to the one used in (Antonacci *et al.*, 1988). Before undergoing syntactical analysis, the relevant phrases occurring within descriptions are isolated by taking into account the coarse semantic trait (activities vs. activity objects) of the nouns involved. After a single description has been analyzed by our grammar, ESR's undergo an intermediate processing in order to reduce morphological and syntactic ambiguity and to take into account ellipses and conjunctions. These phenomena are handled by employing preference schemes improved by semantic control (Hobbs and Bear, 1990).

Finally, ESR's are converted into conceptual relationships by using a many-to-many mapping between syntactic and conceptual relations similar to the one used in (Antonacci *et al.*, 1988); conceptual relationships are subsequently validated by using a semantic knowledge base, and finally merged into a semantic tree. The knowledge representation technique adopted is inspired by the ITL system (Guarino, 1991), where semantic validation reduces to order-sorted unification.

5 Preliminary results

A prototype of the system has been implemented on Macintosh workstation in LPA MacProlog. It has been tested on a significative fragment of our corpus (about 2000 descriptions in the areas of agriculture and services).

The output of the system has been manually tested by a linguist, for correctness and completeness. Errors turn out to be independent of the syntactic algorithm, and are mainly due to (i) lack of semantic knowledge; (ii) lack of lexical knowledge (unknown or ill-formed words); (iii) difficulties with disambiguation and phrase separation.

The query system is now under development.

As far as system efficiency is concerned, the time complexity of the whole analysis process seems to be almost quadratic with respect to the length of the sentence, while the mean understanding time is well below 10 seconds on a Macintosh IIfx.

6 Discussion and conclusions

We would like to make clear that the approach we have presented is seriously limited, by the fact that we are not able to fully exploit syntactic information. This means that complete sentences containing verbs and subordinate clauses cannot be properly handled. However, our experiments show that the approach behaves well in the field of real business descriptions.

Semantic knowledge plays a fundamental role in our system, as it has to validate the syntactic relationships proposed by the shallow parsing algorithm. Currently, our knowledge base consists of a taxonomy of 3000 concepts, together with 360 semantic constraints for the conceptual relations. Although the current prototype uses a hand-written knowledge base, some techniques for semi-automatic extraction of semantic knowledge from our large corpus have been studied in our project (Velardi *et al.*, 1991). The idea is that the non-ambiguous sentence fragments present in the corpus capture specific word usage patterns, which, via a human-controlled generalization process, may generate the kind of semantic constraints we are looking for to put in the knowledge base.

A full syntactic parsing would be too expensive for large corpora, and it would also fail to consider useful information embedded within ill-formed sentences. The encouraging preliminary results indicate that our approach constitutes a good compromise between syntactic completeness and meaning extraction.

In conclusion, the shallow parsing algorithm we have described may play a crucial role in order to trigger a bootstrapping process of knowledge acquisition, giving us some chance to overcome, at least in our domain, a major bottleneck in natural language understanding.

References

Verónica Dahl. Discontinous grammars. *Computational Intelligence*, 5(4):161-179, 1989.

F. Antonacci, M. Russo, M.T. Pazienza, and P. Velardi. Representation and Control Strategies for large Knowledge Domains: An Application to NLP. *Applied Artificial Intelligence*, 2 (3-4), 1988.

Jaime G. Carbonell and Philip J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. *Computational Linguistics*, Special Issue on Ill-Formed Input, 9(3-4), 1983.

Michela Fasolo, Lorenzo Garbuio and Nicola Guarino. Comprensione di descrizioni di attività economico-produttive espresse in linguaggio naturale. In *Proceedings of the 5º Convegno sulla Programmazioni Logica (GULP '90)*, Padova, Italy, 1990.

Jerry R. Hobbs, John Bear. Two Principles of Parse Preference. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Helsinki, Finland, 1990.

Ralph Grishman and Richard Kittredge editors. *Analysing Language in Restricted Domains: Sublanguage Descriptions and Processing*. Lawrence Erlbaum Ass., Hillsdale, New Jersey, 1986.

Nicola Guarino. A Concise Presentation of ITL. To appear in M. Richter editor. *Processing Declarative Knowledge*. Springer-Verlag, 1991.

K. Jensen, G.E. Heidorn, L.A. Miller, and Y. Ravin. Parse Fitting and Prose Fixing: Getting a Hold on Ill-Formedness. *Computational Linguistics*, Special Issue on Ill-Formed Input, 9(3-4), 1983.

Paola Velardi, Maria Teresa Pazienza, and Michela Fasolo. How to encode semantic knowledge: a method for meaning representation and computer aided acquisition. To appear on *Computational Linguistics*, 1991.