A DISTRIBUTED ARCHITECTURE FOR TEXT ANALYSIS IN FRENCH: AN APPLICATION TO COMPLEX LINGUISTIC PHENOMENA PROCESSING

Marie-Hélène STEFANINI and Karine WARREN
Equipe CRISTAL-GRESEC
University Stendhal
BP 25 38040 Grenoble Cedex 9 France
E-mails: stefanini@stendhal.grenet.fr, karine@ull-cristal.grenet.fr

Abstract

Most Natural Language Processing systems use a sequential architecture embodying classical linguistic layers. When one works with a general language and not a sublanguage, there are different cases of ambiguities at different classical levels; and more particularly when one works on complex language phenomena analysis (coordination, ellipsis, negation...) it is difficult to take into account all the different types of these constructions with a general grammar. Indeed, the inconvenience of this approach is the possible risk of a combinatory explosion. So, we have defined the TALISMAN architecture that includes linguistic agents that correspond either to classical levels in linguistics (morphology, syntax, semantic) or to complex language phenomena analysis.

1 Introduction

The goal of this paper is to show that complex linguistic phenomena like coordination, ellipsis or negation, can be defined and processed in an distributed architecture. In the processing of a very large corpus, the problem is to find an approach allowing the best interaction between different knowledge levels (morphological, syntactic, semantic...) in order to reduce the generation of the ambiguities, that occur within any general system of sequential analysis.

Most NLP systems use a sequential architecture embodying classical linguistic layers. Among them one can find systems for English analysis such as ASK [Thomson 85], LOQUI [Binot & al 85], TEAM [Pereira 85] and for French analysis such as SAPHIR [Erli 87] or LEADER [Benoit & al. 86]. Due to the necessity for cooperation between differents modules, we have turned ourselves to the technics of multi-agents systems for the construction of TALISMAN architecture [Stefanini 93]. This system also uses linguistic models of the CRISTAL system [MMI2 89]. The TALISMAN architecture includes linguistic agents that correspond either to classical levels in linguistics (morphology, syntax, semantic) or to complex language phenomena analysis (coordination, ellipsis, negation...).

2 Ambiguities in text analysis

2.1 Examples of ambiguities at different levels in the CRISTAL system:

In NLP, when one works with a general language and not a sublanguage, there are different cases of ambiguities at different classical levels.

Preprocessing: the characters are standardized and the text is cut into forms. So, the punctuations can be ambiguous. For example, a full stop can indicate an abbreviation or the end of the sentence. *M. Clavier* (proper noun/common noun)

Morphology: the text forms are processed individually by the morphological analyser [Aho & Corasick 1975] that attributes one or more interpretations to each in terms of a pair (lexical entry, category).

One of the difficulties is to find the verb in the homonymous sequence with D/Y (determinant/preverbal) F/V (noun/verb). It is possible to predict either the beginning of a noun phrase (SN) or a verbal phrase (SV).

Example: *Pilots* (1) *like* (2) *flying* (3) *planes* (1) *can* (1) *be dangerous.*

(1) *Pilots*, *planes*, *can* are either be verbs (to pilot/to plane) or nouns (a pilot/a plane/a can of beer) (V/F).

(2) *like* is either a verb (to like) or a preposition (like) (V/P). The cooperation between agents in the Talisman system is detailled in [Koning & al 95].

Syntax: A general grammar has rules which interfere with other rules. For example: N"-> N"N" enables to built N" resulting from the concatenation of two N (noun or adjective)". This rule allows to construct the juxtaposition of noun phrases. Example: *Le lyce Louis* (F(nom,ppr) *Le Grand.* But this rule also is applied in the following example: *On associe à* [*chaque étudiant*]SN

[*un numéro de carte*]SN.

Semantics: there are notions of ambiguity and paraphrase. The modals can be paraphrased in a variety of ways. Example: he may come / it's possible that he comes/will come/I permit/authorize/empower him to come.

2.2 Disambiguisation methods

An ambiguity appears when several solutions are possible for the same problem. These ambiguities are produced by a module or are the consequence of different analysis modules.

2.2.1 Local grammars for disambiguisation:

We advocate the use of local grammars for some disambiguisation of several solutions produced by a module. For example, we can use contextual laws for some morphological disambiguisation. Indeed, the following laws are always valid for written french analysis:

- Law 1: *Determiner + (Noun or Verb) -> Determiner(D) + Noun(F)*

English example : "the address ..."

- Law 2: *Pronoun + (Noun or Verb) -> Pronoun(Y) + Verb(V)*

English example: "I address ..."

These laws can be viewed as partial solutions for combinatory explosion.

2.2.2 Interactions for disambiguisation:

In some cases, the interactions between different modules allow a faster disambiguisation. Indeed, an agent can use the knowledge of another agent when needed.

For example, during the morphological and syntactical analysis of the sentence "*I(Y) want(V) the(D) e-mail(F) address(F or V)*", interactions between MORPH and SYNT are useful. MORPH will send all the sure morphological informations to SYNT; MORPH will propose the two morphological interpretations for "*address*". SYNT will immediatly reject the "*address*" = Verb solution in this sentence, thanks to its knowledges.

In other cases, cooperation between agents is needed when two agents produce different solutions for the same problem. For example, the form *and* can be viewed as a syntagm coordinator or as a proposition coordinator.

3 Distributed approach for Natural Language Processing

Blackboards have been applied in linguistics to Speech Understanding Systems [Erman 80] and more recently to the analysis of written French (HELENE [Zweigenbaum 89], CARAMEL [Sabah 90]) and documentary research [Mekaouche 91]. The global control of these systems is fully centralized: the distribution of the reasoning capabilities enforces the maintenance of a global representation that is coherent and thus requires the use of belief revision mechanisms. Architectures based on direct communication between agents allow complete distribution of both knowledge control and distribution of partial results.

We will briefly report on the agent and on agent society concepts as they are defined in [Stefanini 93]. A linguistic agent can be divided into two main parts: its knowledge representation and its knowledge processing. Knowledge and goals can be given or acquired through communication with other agents.

At present, the society in the TALISMAN application is represented by the following linguistic agents: PRET for preprocessing, MORPH for morphological analysis, SEGM for splitting into clauses [Maegaard & Spang Hanssen 78], SYNT for syntactic analysis, TRANSF for transformations of utterances (interrogatives, imperatives, etc...) in declarative clauses, COORD for coordinations, NEGA for negations and ELLIP for ellipses. These agents are described in details in [Stefanini 93]. There are different types of decomposition: Knowledge decomposition by abstraction (PRET, MORPH, SEGM, SYNT...), task decomposition by type of input (COORD, NEGA), task decomposition by type of output (ELLIP).

The TALISMAN system is based on direct communication between agents and thus uses mailboxes for sending messages with an asynchronous mode of communication. Speech acts [Searle 69] are usually used to communicate in a Multi-Agent System. Intentions of the sender are expressed in a common communication language. The possible interactions between agents during a conversation have to be regulated, this is done by means of interaction protocols.

In the TALISMAN system, the communication language and the interaction protocols are based on the work of Sian [Sian 90].

3.1 Messages

In the system, an agent willing to send a message will use the following message format:

((sender, receiver(s)), (performative, force), content).

The name of the sending agent enhances the message understanding and the answer. The sender should determine the addressee agent(s) with the help of its knowledge about the other agents; if he has none, he will send the message to every agent in the system.

The performative of the message is either a simple sending information, a request or a reply. However, these types of messages do not suffice to express all the intentions agents may have. We

have "used" the communication language developed by Sati Sian because it is adapted to the communication needs of the system. This communication language figures out 9 forces : *propose*, *modify*, *assert*, *agrec*, *disagrec*, *noopinion*, *confirm*, *accept* and *withdraw*.

We will not use the force "accept" that requires the agreement of every agents. We also did not use the forces "agreed" and "disagreed" because our agents only have reliable information.

The propositional content is formulated in the knowledge representation language of the agent.

3.2 Communication protocols

An interaction protocol is a set of rules containing the possible interactions during a conversation; it provides strategies for problem solving due to the co-existence of several agents in the same system. For the cooperation between agents, we have adapted the protocol of Sian to the needs of a natural language processing system for written french.

Our protocols will use the language communication defined above. Sian's protocol will be simplified and decomposed for better understanding into three protocols:

- an assertion protocol: this protocol allows agents to send partial or complete results to the concerned agents; it is used when an agent has only one solution or when the work of an agent is finished.
- an information request protocol: this protocol allows an agent to ask a precise question to one or more agents. If the receiver can answer, it will send an "Answer(Assert)", otherwise an "Answer(Noopinion)" (i.e if the agent can not answer or does not understand the question).
- a cooperation request protocol: this protocol allows an agent to ask one or more agents to cooperate with it in order to solve the conflict it has created: it has produced several solutions for the same problem and the other agents have to confirm or reject its hypothesis. An agent will answer noopinion if it can not answer or if it does not understand the question; it will confirm the hypothesis if it obtains a positive evaluation of it and it will withdraw it in case of negative evaluation. If the receiver's agent obtains a negative evaluation and has another hypothesis, it will reply to the sender agent an "answer(modify)" containing its new hypothesis.

Note: when an hypothesis is confirmed and withdraw by different agents, the rejection of the hypothesis will be retained.

4 Example of complex linguistic phenomena processing:

The sentence to process is: "Should (V) I (Y) correct (Fadj/V) the (D) paper (Fnoun/V) and (C(intra / inter)) address (Fnoun/V) him (Y)?"

The process of this interrogative sentence will begin by the sending of the sentence transformed in an affirmative form and preprocessed; The following messages will be sent:

SEND (Pret, Transf; Inform, Assert; [Sentence="Should I correct the paper and address him?"])

SEND (Transf, Broadcast; Inform, Assert; [Sentence="I should correct the paper and address him", QTO])

SEND (Pret, Broadcast; Inform, Assert; [Sentence="I should correct the paper and address him", QTO])

Then Morph agent process the disambiguisation with the linguistic contextual laws presented in the first part. It will find: "I (Y) should (V) correct (V) the (D) paper (Fnoun) and (C(intra / inter)) address (Fnoun,V) him (Y)?" The cooperation between the morphological and syntactical levels can start; Morph send first all the sure informations:

SEND (Morph, Synt; Inform, Assert; ["I=Y","should= V", "correct=V", "the=D", "paper=F", "and=C", "him=Y"])

Then, the coordinator "and" can be viewed by the segmentation (Segm) as a proposition coordinator (inter-proposition coordination) and by the coordination (Coord) as a nominal syntagm (noted SN) coordinator (intra-proposition coordination). But, after the disambiguisaton of "address" the Coord agent will change his point of view. The sending of messages can be done like following:

1- R1 SEND (Morph, (Coord, Segm); Request, /; ["and"=C(intra/inter)])

2- H1 SEND (Morph, Synt; Request, Propose; ["address"=F,V])

3- R2 SEND (Segm, Morph; Request, /; [nb_verb_conjug =?])

4- R2 SEND (Morph, Segm; Answer, Assert; [nb_verb_conjug =2])

5- R1 SEND (Segm, Morph; Answer, Assert; ["and"=C(inter)])

6- I1 SEND (Segm, (Synt, Coord, Ellip); Inform, Assert; ["and"=C(inter)])

7- I2 SEND (Synt, (Coord, Segm); Inform, Assert; ["the paper"=SN])

8- H1 SEND (Synt, Morph; Answer, Assert; ["address"=V,F])

9- I3 SEND (Synt, (Segm, Coord); Inform, Assert; ["address"=V,F])

10- I4 SEND (Ellip, (Synt, Segm); Inform, Assert; [ellips_subject="I"])

11- I5 SEND (Ellip, (Synt, Segm); Inform, Assert; [ellips_c.o.d="The paper"])

12- R1 SEND (Coord, Morph; Answer, Assert; ["and"<>C(intra)])

13- I6 SEND (Coord, (Segm, Ellip, Synt); Inform, Assert; ["and"<>C(intra)])

Legend: Hi : is the hypothesis i on which the agents have to work.

Ri : is the information request i at which the agents have to answer.

Ii : is the information sending i.

In fact, this is an example of a possible development of the interaction protocols by the agents concerned by the coordination phenomena. But the use of pseudo-parallelism and asynchronous sending of messages can provide different sending of messages.

5 Conclusion

In this paper, we have proposed a method to solve some ambiguities and some complex linguistic phenomena in a TALISMAN Multi-Agent System. To allow cooperation and resolution of conflicts, we have developed interaction protocols adapted to the needs of a natural language processing system for written french. These interaction protocols allow cooperation and resolution of conflicts that appear at one time in the system, particularly during complex linguistic phenomena treatment.

Currently, we are integrating the prototype linguistic agents (which implement different types of coordination, negation and ellipsis) in order to validate the developed protocols. The implementation is realized with Prolog II+ on an IBM Workstation. After the implementation, we will be able to evaluate and possibly refine the cooperation and conflicts resolution methods that have been developed.

References

Benoit P., Rincel P., Sabatier P., Vienne D. A user-friendly natural language interface to oracle. In Proccedings of the European Oracle Users' group Conference, Paris, 1988.

Berrendonner A., "Grammaire pour un analyseur, aspects morphologiques", Les Cahiers du CRISS, N 15, Novembre 1990.

Binot J.L., Demoen B., Hanne K., Solomon L., Vasiliou Y., von Hahn W., and Wachtel T. LOQUI: a logic oriented approach to data and knowledge bases supporting natural language interaction. In Proceedings of ESPRIT 88 Conference, 1988.

Aho A. V., Corasick, M.J., " Efficient string-matching: An aid to bibliographic search", Communications of the ACM, Vol. 18, N6, June 1975.

Erli. SAPHIR : manuel de description du logiciel. Technical report, Socit rli, 1887.

Esprit Project 2474. MMI2. Common meaning representation by D. Sedlock. Report Bim/13, October, 1989.

Koning J.L., Stefanini M.H., Demazeau Y. "DAI Interaction protocols as Control Strategies in a Natural Language Processing System". Proceedings of IEEE International Conference on Systems, man and Cybernetics, Vancouver, October, 1995.

Maegaard B., & Spang Hanssen E., "Segmentation automatique du français écrit", documents de linguistique quantitative, Dunod 1978.

A. Mekaouche, J.C. Bassano,Analyseur linguistique multi-expert pour la recherche documentaire. Avignon, Mai 1991.

F. Pereira, "The TEAM Natural-Language Interface System", Final Report, SRI International, Menlo Park, 1985.

Thomson B., Thompson F. : "ASK is Transportable in Half a Dozen Ways". ACM Transactions on Office Information Systems, Vol 3, N 2, April 1985.

Searle J.R., "Speech Acts" by Cambridge University Press, 1969.

Sian S.S., Adaptation based on cooperative learning in multi-agents systems". In Proceedings of the European workshop on Modelling Autonomous Agent in a Multi-Agent World, MAAMAW 1990.

M.H. Stéfanini TALISMAN : un système multi-agent pour l'analyse du français écrit, Thèse de Doctorat, Jan. 1993.

P. Zweigenbaum, Helene: Compréhension de compte-rendus d'hospitalisation. Deuxième Ecole d'été sur le Traitement des Langues Naturelles. L'ENSSAT, Lannion, Juillet 1989.