Distributing linguistic knowledge in a multi-agent natural language processing system: re-modelling the dictionary

Vera Lucia Strube de Lima
PUCRS - Informatica (Brazil), MAGMA-LEIBNIZ (France)
vera@andros.inf.pucrs.br

Paulo Ricardo Carneiro Abrahão Ivan Santa Maria Filho
PUCRS - Informatica (Brazil)
abrahao@andros.inf.pucrs.br, ifilho@cpovo.net

Abstract

Most Natural Language Processing (NLP) systems traditionally use a sequential architecture that represents the classical linguistic levels (lexical-morphological, syntactical, semantic, etc), but they claim for huge and monolithic modules, that make the backtracking among such modules difficult. By using a distributed architecture, we take advantages such as the parallelism in the procedural levels of the system, the distribution of the memory resources like knowledge and methods, and the interaction among the modules, wich can change information by message passing. This article puts emphasis on a distributed proposal for the dictionary structure, in a NLP distributed environment. This proposal is oriented through a distributed perception of language processing, in a multi-agent system. Knowledge associated to morphology, syntax and semantics is embedded in specific agents, which are members of a NLP agents society. We present the dictionary models used for each of those agents, and we show how they are integrated in a system which processes Portuguese sentences.

1. Introduction

The multiple forms of linguistic knowledge have to be represented in a NLP application, so that we can deal with semantic and pragmatic flexibility, with the structural richness of natural language, and with complex linguistic phenomena such as ambiguity or reference. To deal with this complexly related information, and to make it available for text analysis, some previous studies [Fum 88, Fuchs 93, Stefanini 92, Csuhaj-Varjú 93, Boitet 94, Guha 94, Silva 97, Lima 97] have pointed to a distributed architecture, opposite to a sequential one. Some of them, as [Fum 88, Stefanini 92, Silva 97, Lima 97] report research on the possibilities of using a social metaphor (the MAS or, multi-agent systems [Demazeau 90]) in NLP, to represent cooperation among distinct linguistic levels.

An agent can be introduced [Boissier 91] as an intelligent entity able to act and understand a known environment, pursuing its aims in a rational and intentional way, according to the current state of its knowledge (its internal state), that changes by cooperation with other agents in the same environment.

The aim of our linguistic agents is to participate in a society of entities with different skills, and to collaborate in the interpretation of natural language (esp. Portuguese) sentences.

Distribution and cooperation among agents mean to deal not only with methods, but also with knowledge associated with NLP systems and, the first knowledge source to consider is, undoubtly, the dictionary.

Centering on representation and implementation of Portuguese dictionaries for a NLP system, [Kowaltowsky 95, Nunes 96] propose very performant alternatives while they use binary automata to represent lexical items. This kind of representation, however, has to be enriched with syntactical and semantic knowledge in order to be useful for natural language applications such as sentence interpretation. [Dias 94] approaches not only lexical but also syntactical and semantic information, in an organization inspired from the ones of Jackendoff and Barsalou. However,

the work of Dias can be re-thought, regarding more recent models such as the Generative Lexical Semantics [Pustejovsky 95] and, mainly, according to a distributed approach.

We have experimented a distributed alternative for a NLP system which processes Portuguese sentences, so that its dictionary was split among the different agents involved. In this paper we detail the organization adopted in our experiment for the dictionary information in this distributed environment.

2. A distributed approach

2.1 General Organization

Morphological, syntactical and semantic information about the lexical items are necessary for the resolution of specific tasks, during language analysis. For example, information about the transitivity of a verb, as well as about its complements, is interesting for a syntactical analysis. Semantic information such as the knowledge necessary for reference resolution (like pronominal anaphora resolution) is sometimes unnecessary for a syntactical analysis, although this is very important for comprehension.

These evidences let us experiment a society composed of two kinds of agents:

- the agents associated with linguistic levels (as 'regular' agents according to [Stefanini 92]);
- agents associated with the resolution of linguistic phenomena as anaphora, ellipsis, etc (in some aspects, like "transversal" agents according to [Stefanini 92]).

Agents from the first group seem to be directly related to the dictionary, so that dictionary knowledge is distributed among them. In the following sections, we present the dictionary perspective from each of those agents.

In order to carry on this experiment, our lexical items are stored in a TRIE[1] tree structure. From its root, we access the 26 letters of the alphabet, that branch into strings of characters. A string may represent a complete lexical item (so it is a *form*, e.g., the complete expression 'a priori') to which we attach descriptors. Or, a string may represent the first part of an entry, called *base* (e. g., the string 'cant' for the verb 'cantar'/*to sing* in English), to which we also attach descriptors.

[1] A special data structure whose basic idea is to compare entry string parts.

2.2 A morphological agent

The morphological dictionary is represented by a structure that allows entry decomposition in basic components which are bound to morphological information groups or, *models*. Items can be stored as forms or as bases, depending on the linguistic connotation to be given to a certain item. While form entries are stored as complete strings of characters, base entries are stored as an invariable part (the base) that points to one or several models, which associate this base with terminations.

The descriptors attached to bases point to *models*, and the models contain references to the possible *terminations* for a base. The base, attached to its termination, shall constitute a valid Portuguese lexical item (word or expression). Besides referencing to a termination, models contain information on morphological variables as number, gender etc.

A string of characters that constitutes a form or a base is associated with one or more information registers, called *descriptors*. The descriptor stores lexical-morphological information about the entry and, if this entry is a base, it points to a model. An entry will have as many descriptors as the different types it represents. For example, the base *cas* has two descriptors - one for the model that builds the feminine noun *casa* (*house* in English) and its inflexions or derivates (plural *casas*, diminutive *casinhas* etc), another one for the model that builds the verb *casar* (*to marry* in English) and its conjugated forms (*caso*, *casas*, ..., *casaria*, ..., *case*, *cases*, ...).

In Portuguese language, to groups of common nouns for example, there are standard rules for changes in gender, number, diminutive/augmentative etc. With bases and models, it is possible to represent these changes by binding the base to the appropriate terminations, using different models. Figure 2.2.1 shows a model for feminine nouns ending on *a*.

Models allow binding termination groups to bases performing identical inflexion behavior, achieving reduction in the number of both characters and descriptors stored in the dictionary. For example, for the regular verb *cantar* (to *sing*, with 56 inflexions in

Portuguese) it is necessary only one descriptor, that points to the model that produces all its inflexions. With models, variable information (gender, number, degree, tense, aspect, ...) is stored with terminations (just once), avoiding information redundancy.

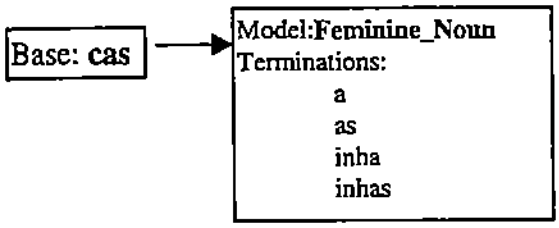

FIGURE 2.2.1 FEMININE NOUN MODEL

As a morphological analyser, we provide a component that scans sentences and goes through the TRIE tree structure looking for a string or for the first part of a string matching to that sentence or to a part of it. This component is able to accept a sentence written in Portuguese and to split it into lexical items, returning, for each item, the set of different types (and morphological variables associated) it can represent. This component doesn't deal, itself, with lexical ambiguity.

Our morphological entity, as an agent, can communicate either with applications or with other agents in the society. However, it works independently from the other agents.

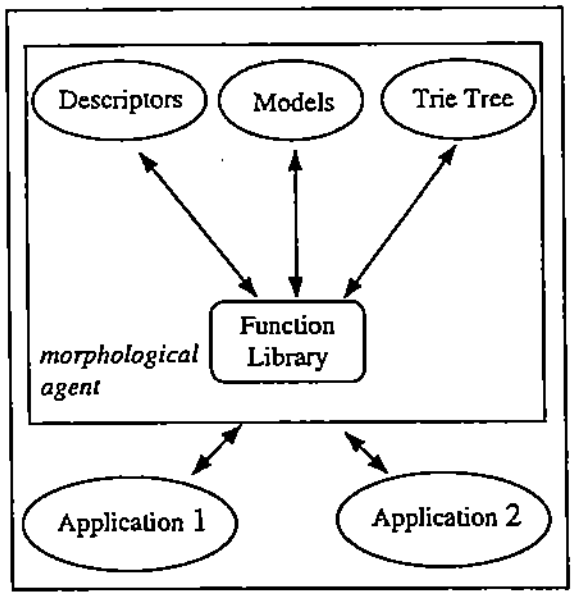

FIGURE 2.2.2 AN AGENT-BASED MODEL FOR MORPHOLOGY

The morphological dictionary works as a function library on a structure set. This library was built to provide morphological information to be shared among several applications, or to allow an application to work as an agent in a NLP multi-agent system. In this case, the library and its structures would constitute the information agent core (see figure 2.2.2) and an interface is responsible for communication with any other agents or applications.

2.3 A syntactical agent

We chose the Tree-Adjoining Grammars (TAG) with lexicalization as the formalism to be used in our experiment. It provides means to connect the lexical items to grammar structures. In the following paragraphs we briefly present TAG, stressing aspects of its lexicalization inside the syntactical agent.

TAG was introduced by Aravind Joshi et al. in 1975, as a formalism for linguistic description. A TAG's basic component is a finite set of elementary trees, each of which is a domain of locality, and can be viewed as a minimal linguistic structure. A TAG comprises of two kinds of elementary trees: initial trees, which are complete structures, and auxiliary trees, that have at least one leaf symbol of the same category of the root symbol. Sentences are derived from the composition of the initial tree and any number of auxiliary trees by an operation called 'adjunction'. Adjunction inserts an auxiliary tree at one of the corresponding nodes of an elementary or a derived tree and moves the original root to the leaf of the same category. Recursion is provided by the auxiliary trees which can adjoin to themselves. Features, such as prepositions in verbal complements, can be associated with each node of an elementary tree.

The Lexicalized Tree-Adjoining Grammar (LTAG) [Joshi 85] is a way of associating with the dictionary the concept of TAG. According to LTAG, an entry in the lexicon anchors a group of trees where it can appear. This approach avoids a lot of backtracking in the recognition process, but brings a new set of problems. The grammar must be suitable to the context of text and must be intimately associated with the lexicalization dictionary.

In order to implement this model, grammar knowledge comprising the initial tree models, the auxiliary tree models that represent the structure of Portuguese sentences and the lexicalization dictionary, is part of the syntactical agent knowledge. The syntactical

agent itself can be seen as a subsociety, formed by agents handling simpler tasks, so that the syntactical knowledge is itself distributed among those simpler components. Information associated with the features (e.g. complements) of verbs is organized with the trees (which gather information useful for parsing). This subsociety can be dinamically organized, according to the problem it is expected to solve.

One of the possible organizations for this subsociety is:

- agent 1, handling auxiliary trees;
- agent 2, handling initial trees;
- agent 3, handling the lexicalization dictionary and,
- agent 4, handling the formalism operations and organizing the internal memory of the subsociety.

The initial trees agent as its counterpart, the auxiliary trees agent, handles the grammatical information like a source of structures. The intelligence here involves the selection of the structures to be hidden from the rest of the subsociety, and the ones to be shown. This decision can be taken in contact with "transversal" agents (section 2.1) that handle complex linguistic phenomena such as anaphora or ellipsis, context, or with other agents remarkably the semantic agent.

The lexicalization dictionary agent must negotiate with agents 1 to 3 after getting morphological information (from the morphological agent). It sends to agent 4 the set of trees that must be operated, the restrictions to operations (e.g. Portuguese verbs demands specific prepositions in special situations) and the sentence token that anchors the information. Agent 4 tests all possible combinations with the received information, sets values of the internal memory and sends status messages to the semantic agent.

2.4 A semantic agent

Semantic representation of lexical items (lexical semantics) or, semantic information associated to lexical items, is fundamental to language interpretation, and it allows constructing knowledge representation for sentences or even texts.

The model proposed by Dias [Dias 94] incorporates semantic knowledge associated to words, but the composition of lexical items (using Jackendoff's and Barsalou's model) is strict. Studies by James Pustejovsky [Pustejovsky 95] propose a theory for the representation of semantic knowledge in a generative lexicon. The generative lexicon is characterized as a computational system involving at least four levels of representation: argument structure, event structure, *qualia* structure and lexical inheritance structure. Argument structure indicates the way a certain lexical item is to be associated to a syntactical expression. Event structure identifies a particular event type to a lexical item or expression. Qualia structure has the essential attributes for an object, event and relations that define the lexical item. Inheritance structure establishes the relationship among a lexical item and other concepts in the lexicon.

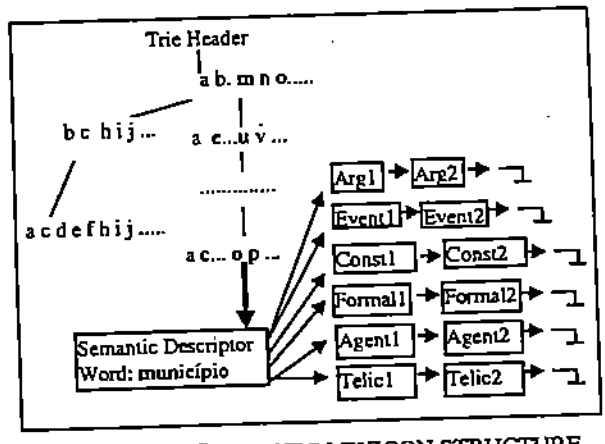

FIGURE 2.4.1 SEMANTIC LEXICON STRUCTURE

The elements that build Qualia structure include notions of the constituents, space, figure, surface, manufacture etc, and are organised in four aspects of the meaning of a lexical item, each aspect refering to a role associated to meaning. Those roles are known by the denominations *constituent*, *formal*, *telic* and *agent*.

Following to the organization proposed by Pustejovsky, we have implemented the semantic lexicon from a TRIE which stores the lexical items. Semantic information associated to the items is stored in chained lists, from a structure called semantic descriptor (see figure 2.4.1).

Implementation was based on three main structures: arguments, events and Qualia. The two first ones use each one a list, and the last one is implemented using four different lists (one for each role in Pustejovsky's model).

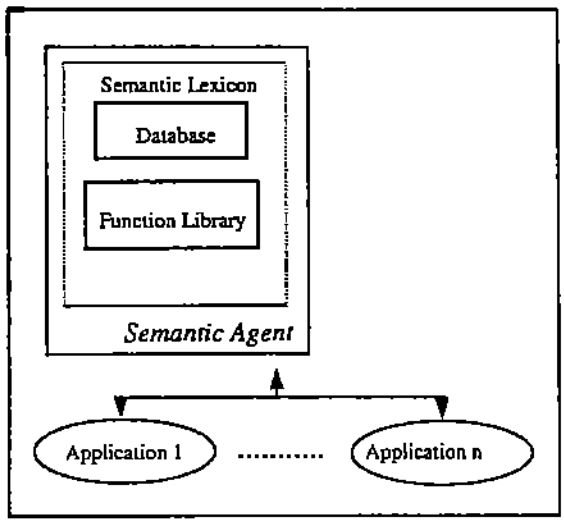

FIGURE 2.4.2 THE SEMANTIC LEXICON IN MAS

The implementation proposed for the semantic lexicon was conceived to be part of a multi-agent system. This lexicon works as a function library and acts on a lexical database[2]. Figure 2.4.2 shows situates the semantic lexicon in a multi-agent system.

The semantic agent provides semantic information about lexical items by means of the operations (e.g. type coercion) proposed by Pustejovsky. This agent is called to collaborate in solving linguistic phenomena such as pronominal anaphora. It provides, then, the onthology for nouns etc. It also participates in ambiguity solving. Nevertheless, it is potentially useful in many other applications as a self-contained device.

3. Conclusions

This article presents a distributed approach to the lexicon in a NLP system, putting emphasys on dictionary distribution. Our main agents in this experiment (see figure 3.1), have been attached to morphology, syntax and semantics, and dictionary knowledge has been distributed among those three agents.

In order to deal with specific phenomena such as ambiguity, we have given those regular agents the capacity to dinamically form subsocieties (in figure 3.1, shown as triangles). For example, a lexical categorial ambiguity is detected when a lexical item has two or more descriptors connected to it. Ambiguity solving is a sort of conflict solving, with the cooperation of the syntactical and the semantic agents.

Other phenomena such as anaphora resolution, are treated by extra entities or, specific transversal agents, even though the dictionary knowledge was kept under the three main agents.

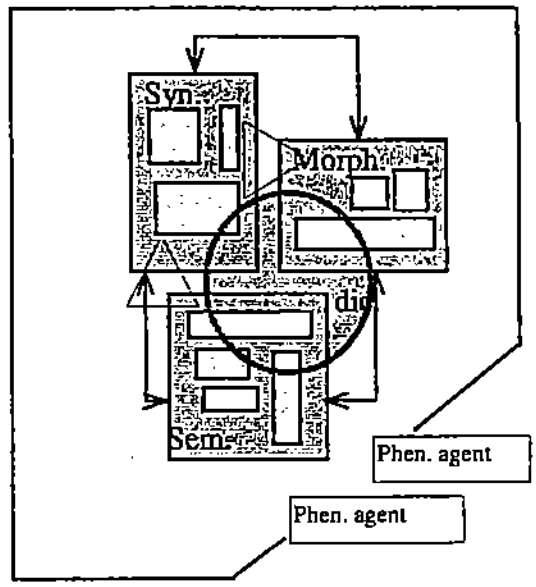

FIGURE 3.1 NLP MAS ORGANIZATION EXPERIMENTED

Our approach has inherited modular and serial characteristics from more traditional Linguistics. It would be possible, however, to differently conceive agents' constitution and even cooperation among agents. Works as [Paiva 96] propose word-based agents for Portuguese language interpretation, which show higher granularity but, on the other hand, are supposed to be internally simpler than the 'linguistic level' -based agents.

A multi-agent perspective on language processing has also to be considered as a kind of migration, from some established work, to a new architecture. Even if components were prepared to collaborate with other components in a NLP system, as agents, the organization (and architecture) of those components (with breakpoints where to promote communication and exchanges) has to be re-thought before they effectively collaborate in solving problems.

We consider that a distributed proposal better approaches NLP problems, once it allows the use of different formalisms to each level of linguistic processing, it allows handling with different knowledge sources in a modular way, and it provides storage

[2] According to [WILKS96], a lexical database is a structure storing lexical information which can be used by programs that process language.

distribution for the lexicon. In future experiments, we intend to re-organize this first architecture and to extend it for a larger number of linguistic phenomena.

This experiment was carried on under the NALAMAS[3] Project, which investigates multi-agent architectures for natural language processing systems, namely for Portuguese language.

4.References

[Boissier 91]

Boissier, O. & Demazeau, Y. (1991). **A distributed artificial intelligence view on general purpose vision systems.** Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Kaiserslautern, Germany, August 5-7.

[Boitet 94]

Boitet, C. & Seligman, M. (1994). **The Whiteboard Architecture: A way to Integrate Heterogeneous Components of NLP Systems.** Proceedings of COLING-94, Vol. I, Kyoto, Japan, August 5-9.

[Csuhaj-Varj 93]

Csuhaj-Varj, E. & Alez, R. A. (1993). **Multi-Agent Systems in Natural Language Processing.** In K. Sikkel, A. Nijholt eds. Twent Workshop on Language Technology 6, pg. 129-137, December 16-17.

[Demazeau 90]

Demazeau, Y. & Muller, J. P. (1990) **Decentralized AI.** Morgan Kaufmann.

[Dias 94]

Dias, M. C. P. (1994). **O léxico em sistemas de análise e geração automática de textos em língua portuguesa.** Ph.D. Thesis, PUCRJ, Rio de Janeiro.

[Fuchs 93]

Fuchs, C. (1993). **Linguistique et Traitement Automatique des Langues.** Hachette Université, Paris.

[Fum 88]

Fum, D., Guida, G. & Tasso, C. (1988). **A Distributed Multi-Agent Architecture for Natural Language Processing.** Proceedings of COLING-88, Vol. II, Budapest, August 22-27.

[Guha 94]

Guha, R. V. & Lenat, D. B. (1994). **Enabling Agents to Work Together.** Communications of the ACM, (37) 7: 127-142.

[Joshi 85]

Joshi, A. (1985). Tree-Adjoining Grammars: how much context-sensitivity is required to provide reasonable descriptions? Editors Dowty, D., Kartunen, L., Zwicky, A. **Natural Language Parsing: psychological, computational and theoretical perspectives.** Cambridge University Press, Cambridge MA.

[Kowaltowski 95]

Kowaltowski, T., Lucchesi, C.L. & Stolfi, J. (1995). **Minimization of Binary Automata.** Journal of the Brazilian Computer Society, nro. 3, vol. 1, April.

[Lima 97]

Lima, V. L. S. et alii (1997). **Uso de sistemas multi-agentes no processamento de linguagem natural: estudando o problema através do projeto NALAMAS.** Proceedings of the ENIA'97. Brasília, August 3-4.

[Nunes 96]

Nunes, M.G.V. et alii. (1996). **A construção de um Léxico para o Português do Brasil: lições aprendidas e perspectivas.** In: Anais do XIII Simpósio Brasileiro de Inteligência Artificial SBIA96. Curitiba.

[Pustejovsky 95]

Pustejovsky, J. (1995). **The Generative Lexicon.** MIT Press, Cambridge.

[Silva 97]

Silva, J. L. T , & Lima, V. L. S. (1997). **An alternative approach to lexical categorical desambiguation using a multi-agent systems architecture.** Proceedings of the RANLP'97, Bulgaria, September 6-12.

[Small 87]

Small, S. L. (1987). A distributed word-based approach to parsing. Editor Leonard Bolc in **Natural Language Parsing Systems,** Springer-Verlag, Berlin.

[Stefanini 92]

Stefanini, M. H. , Berrendoner, A., Lallich, G. & Oquendo, F. (1992). **TALISMAN: un système multi-agents gouverné par des lois linguistiques pour le traitement de la langue naturelle.** Proceedings of the COLING'92, Nantes, August 23-28.

[Wilks 96]

Wilks, Y.A., Slator, B. M. & Guthrie, L. M. (1996). **Electric Words.** The MIT Press, Cambridge, Massachusetts.

[3] Funds from the Brazilian Agency CNPq/PROTEM-CC #680081/95-0