TRANSFORMING ENGLISH INTERFACES TO OTHER NATURAL LANGUAGES: AN EXPERIMENT WITH PORTUGUESE

GABRIEL PEREIRA LOPES (1)
Departamento de Matemática
Instituto Superior de Agronomia
Tapada da Ajuda - 1399 Lisboa Codex, Portugal

ABSTRACT

Nowadays it is common the construction of English understanding systems (interfaces) that sooner or later one has to re-use, adapting and converting them to other natural languages. This is not an easy task and in many cases the arisen problems are quite complex. In this paper an experiment that was accomplished for Portuguese language is reported and some conclusions are explicitely stated. A knowledge information processing system, known as SSIPA, with natural language comprehension capabilities that interacts with users in Portuguese through a Portuguese interface, LUSO, was built. Logic was used as a mental aid and as a practical tool.

1. INTRODUCTION

The CHAT-80 program for English (Warren & Pereira, 1981; Pereira, 1983) was transformed and adapted to Portuguese. Logic Programming as a mental aid, and Prolog (Coelho, 1983; Clocksin & Melish , 1981) and Extraposition Grammars (Pereira, 1983) as practical tools, were adopted to implement a natural language interface for Portuguese. The interface here reported, called LUSO, was then coupled to a knowledge base for geography, an extension of the CHAT-80 knowledge base. In an ulterior experiment , LUSO dictionary was augmented with new vocabulary and LUSO was coupled to other modules that considerably augmented the expertise capabilities of SSIPA (Sistema Simulador de um Interlocutor Português Automático (2)).
SSIPA is a complex knowledge information processing system with natural language comprehension and synthesis capabilitites that interacts with users in Portuguese due to the linguistic knowledge that is logically organized and codified in the above mentioned SSIPA's interface called LUSO. After the first step of its development, SSIPA was able to answer questions about geography and could agree or disagree with the opinions stated by the users about its geographical knowledge. After the second step of its development SSIPA became more powerful and intelligent because it could also perform actions that traditionally were attributes of computer monitors (Lopes & Viccari, 1984). As a matter of fact, SSIPA can create and delete files, fill them, change their names, list and change their contents; SSIPA receives, keeps and send messages answers questions not only about geography but also about the knowledge SSIPA represents; it agrees or disagrees with the opinions stated by users about the Knowledge context behind dialogues, reacts when users try to cheat it but, as a rule, SSIPA behaves as a helpful, deligent and cooperative interlocutor willing to serve human users, changing from one to another topic of conversation and developing intelligent clarification dialogues (Lopes, 1984). All these features require a very powerful Portuguese language interface whose main morpho - syntactic features are pointed out in this paper.

2. FORMALIZATION OF NATURAL LANGUAGE CONSTRUCTS

Natural language are complex structured systems difficult to formalize. Formalization can be understood as a step by step construction of a theory to achieve , as an ultimate goal, an axiomatic definition of natural language constructs. If this descriptive theory can also function as the linguistic structured knowledge necessary to simulate a human native using his mother language then, the formalization effort has acquired and gained a new insight. While representing a natural language system, it may represent a native competence about his mother language and, simultaneously, it may perform the role of a native using that competence. This dual unity, incorporating a description of linguistic knowledge and incorporating the same linguistic knowledge ready to be active, is central to this work. This unification in the same unit of two apparently conflicting and contraditory aspects of natural languages is possible due to the usage of logic as a mental and a practical tool. SSIPA encapsulates both views of natural language.

Practice demonstrates that, for the construction of complex models it is better to begin with simple model versions to represent the system one intends to simulate. This practical conclusion

---

(1) Present Adress: Centro de Informática, Laboratório Nacional de Engenharia Civil, 101, Av. do Brasil, 1799 Lisboa Codex, Portugal

(2) Simulating System of a Portuguese Automatic Interlocutor.

seems reasonable because knowledge about a system and about its representation keeps on augmenting as far as, to achieve the validation of the simula - ting model, empirical investigation progresses(Klir, 1975). However one must be aware that while Know - ledge about a real system keeps on growing so do the complexity that one can unwillingly introduce in to the model. Having all this in mind, if we want to formalize linguistic knowledge about natural lan guage we must be prepared to use powerful formal languages prone to description of complex systems and able to be used as programming languages. Here it is subsumed that computers are tools adapted to deal with complexity, augmenting considerably human capabilities to handle highly complex represen tational systems.

3. LUSO

LUSO input subsystem is a device that transforms a sequence of words morfologically, syn tactically and semantically significant into a Logical Form. A Logical Form is here understood as a sequence of predicates, envelopes for knowledge transportation from users to SSIPA central processing unit (the EVENT DRIVER) and from this unit to users. These predicates generalize and augment the potencialities of Pereira's equivalent predicates, (Pereira, 1983). They can also be compared with the lexical functions of Bresnam (1981). However we don't use case classification. In Portuguese, prepositions associated to noun semantic features seem to be enough to identify and differentiate meanings of verbal, noun, adjectival and even prepositional form functions (Lopes, 1984).

LUSO is a natural language interface that concentrates linguistic expert knowledge about Portuguese language.

LUSO input subsystem works sequentially. In a first step it performs the syntactical analysis of an input Portuguese sequence of words. Depending on the task LUSO has been commited to perform, a lexically filled syntagmatic marker or a failure is the result of LUSO eagerness to prove the above mentioned input sequence of words as a syntactically correct yes-no question, wh-question, imperative or declarative sentence, or as a syntactically correct noun phrase or prepositional phrase. When a lexically filled syntagmatic marker is obtained, it is translated to a logical form. Finally this form is planned and simplified accor - ding to the methodology described by Pereira (1983) and Warren (1981).

The design of LUSO input subsystem re - flects the following hypothesis:

- morphological analysis of Portuguese constructs is syntactically driven;
- linguistic semantic analysis of Portuguese constructs is lexically (functionally) driven (in a quasi-bresnamian, sense (Bresnam, 1981; Pereira, 1983;Lopes, 1984));
- cognitive semantic analysis of Portu - guese constructs depends on syntacti - cal and linguistic semantic analysis previously achieved for Portuguese constructs.

This suggests SSIPA as a formal system that already theorizes some aspects of Portuguese language while LUSO specificates the form of formal functions whose cognitive content and formal aptitude for transforming system state are defined at the semantic level of the formal system.

To complete the formal role we wanted SSIPA to play, LUSO output subsystem synthesizes Portuguese noun phrases, prepositional phrases or sentences whenever it receives correspondent requests to output such constructs. To achieve that goal LUSO transforms any previously lexically filled syntagmatic marker into a sequence of Portuguese words in its final forms, ready to be sent to a user.

4. MORPHO-SYNTACTICAL ANALYSIS AND SYNTHE - SIS OF PORTUGUESE LANGUAGE CONSTRUCTS

The morpho-syntactical analysis of Portuguese language constructs is application indepen - dent and is based on the various concepts developed by Chomsky and followers in the framework of the Extended Standard Theory of Generative Grammar (Chomsky, 1980, 1981a, 1981b; Rouveret, 1983 and many others). As it was already mentioned in this paper, one of the crucial hypothesis behind LUSO's design reflects the idea that morphological analysis of Portuguese constructs is syntactically driven. This means that when the syntactical parser is waiting for a specific grammatical category, it takes the next word to be analysed from the input sequence of words and searches the dictionary for that category, trying to find the input word. If the input word does not match any dictionary entry for that particular category, all possible input word endings, one after another, starting from the longest towards ths shortest, are matched against the ending entries for that category until a success - ful match will occur. If such a match does not succeed, this means that the input word does not belong to the foreseen grammatical category. As a consequence, a failure occurs and the Prolog mecha - nism for backtracking is automatically activated. When one of the input word possible endings mat - ches an ending entry for the syntactically predicted category, a basic form for the input word is coined. The newly coined basic form for that in - put word is then checked against the subdictionary entries for the foreseen grammatical category. A process of successes and/or failures proceeds. A syntagmatic marker for each input Portuguese construct is filled with word basic forms and correspon - ding syntactic features information (person, gender and number for noun phrases; tense, mode, aspect , voice and negation for verbs; etc.). The basic form for a verb is its infinitive form; for a noun is its singular form; for a pronoun, article or adjective is its singular masculine form.

The morphological synthesis of Portuguese constructs is syntactically driven. This means that, departing from a syntagmatic marker lexically filled with basic forms of Portuguese words, using the syntactic features that are explicitelly considered into that marker, LUSO output subsystem coines the corresponding sequence of Portuguese words in its final output form ready to be sent to the user with whom the system is interacting. For this purpose most of the rules that were designed to consult LUSO's dictionary were reordered. Departing from basic forms of words, their final forms are obtained by a process nearly inverse of the process used for input.

Extraposition grammars, the formalism developed by Pereira (1983), were used to implement the analyser and the synthesizer for Portuguese. It is worth telling that this formalism proved to be quite adequate for the description of move-alpha rule (Chomsky, 1981b) in complex syntactical environments such as those that frequently occur in Portuguese. As a matter of fact phrase constituents order in Portuguese sentences is quite free. LUSO takes into account the same type of problems handled by CHAT-80 program. Additionally, it analysis syntactical structures involving prepositional phrases and verb headed sentences where there is reordering of noun phrase constituents inside those sentences due to the heading process. Problems related to common nouns followed by the proper nouns they refer, in the context where they appear, is also handled.

5. CONCLUSIONS

It is wiser to concentrate efforts to obtain more and more powerful morpho-syntactic analysers, linguistic semantic analysers and cognitive, semantic interpreters for the natural language we are working in. Constructing replicants of application directed interfaces starting from scratch is unproductive. Constructing more and more powerful interfaces, as the number of applications naturally grows, the natural language analyser, planned to be application independent, is always under improvement because it is always incorporating more and more linguistic knowledge. At the same time one is freed from consideration of morphological and syntactic basic problems and so one can shift his attention to more subtle problems related to tense, modality and others and one can concentrate his mind to the way how concepts related to words are defined. As a consequence, the implementing task can be organized by areas of specialization.

When one has to construct an interface for a specific language it is reasonable to look for interfaces implemented for other languages where the faced syntactical and morphological problems have a similar degree of complexity. Having this in mind, Portuguese language seriously competes with English because it rises quite important syntactic, semantic and pragmatic problems similar to problems risen by latin, slavonic and germanic languages.

6. AKNOWLEDGEMENTS

I would like to thank Helder Coelho for his insightful comments and suggestions throughout this research and the writing of this paper.

7 REFERENCES

BRESNAM, J., "The passive in lexical theory", Occasional Paper 7, The Center for Cognitive Science MIT, 1981.

CHOMSKY, N., "On binding", Linguistic Inquiry, vol. 11, nº 1, 1-46, 1980.

CHOMSKY, N., "Lectures on government and binding", Foris Publications, Dordrecht, Holland, 1981a.

CHOMSKY, N., "On the representation of form and function", The Linguistic Review, vol. 1, nº 1, 30-40, 1981a.

COELHO, H., "The art of knowledge engineering with Prolog", INFOLOG PROJ, Faculdade de Ciências, Universidade Clássica de Lisboa, 1983.

KLIR, G., "On the representation of activity arrays", Int. J. General Systems, 2, 149-168, 1975

LOPES, G., "Implementing dialogues in a knowledge information system", paper submited to International Workshop on Natural Language Understanding and Logic Programming, Rennes, France, 1984.

LOPES, G. and VICCARI, R., "An intelligent monitor interacting in Portuguese language", short paper accepted for ECAI-84, Pisa.

PEREIRA, F., "Logic for natural language analysis", Technical Note 275, SRI International, 1983.

ROUVERET, A., unpublished lectures lectured in Lisbon, 1983.

WARREN, D., "Efficient processing of interactive relational data base queries expressed in logic", Dept. of Artificial Intelligence, Univ. of Edinburgh, 1981.

WARREN, D. and PEREIRA, F., "An efficient easilly adaptable system for interpreting natural language queries", DAI research paper nº 155, Univ. of Edinburgh, 1981.