Possessive Pronominal Anaphor Resolution in Portuguese Written Texts

Ivandré Paraboni, Vera Lúcia Strube de Lima

PUCRS - Instituto de Informática
Av. Ipiranga, 6681 prédio 16
90619-900 - Porto Alegre RS - Brasil
phone # 55 51 320-3500 /3611 Fax # 55 51 3203621
paraboni@tca.com.br, vera@andros.inf.pucrs.br

Abstract

This paper describes a proposal for Portuguese possessive pronominal anaphor (PPA) resolution, a problem little considered so far. Particularly, we address the problem of Portuguese 3rd person intrasentential PPAs *seu/sua/seus/suas* (his/her/their/its, for human and non-human subjects in English), which constitute 30% of pronominal occurrences in our corpus (Brazilian laws about environment protection). Considering some differences between PPAs and other kinds of anaphors, such as personal or demonstrative pronouns, we define three knowledge sources (KSs) for PPA resolution: surface patterns (taking in account factors such as syntactic parallelism), possessive relationship rules and sentence centering. These knowledge sources are organized in a blackboard architecture for PPA resolution, which provides both knowledge and procedure distribution among autonomous entities (reflexive agents), each of them specialized in a particular aspect of the problem solving. The proposal has been implemented and its results are discussed at the end of the work.

1. Introduction

Most work on anaphor resolution apply syntactic constraints (c-command, gender and number agreement, etc) to select the appropriate anaphoric referent. However, these constraints are not suitable for possessive pronominal anaphor (PPA) resolution in Portuguese, which requires a more specific approach.

This paper describes a resolution strategy for a problem little considered so far, PPAs "seu/sua/seus/suas" (his/her/their/its, for human and non-human subjects in English), which constitute 30% of pronominal occurrences in our corpus (Brazilian laws about environment protection).

The paper is structured as follows. We present some characteristics of Portuguese PPAs (section 2). We then describe some factors in PPA resolution and the way these factors can determine PPAs antecedents (section 3). Factors are implemented as knowledge sources in a blackboard architecture (section 4), and finally we present the results of our implementation (section 5).

2. The PPA resolution problem

From the interpretation point of view, PPAs are widely different from other kinds of anaphors, such as personal or demonstrative pronouns. In this section we present some specific characteristics of Portuguese PPAs "seu/sua/seus/suas", by means of generic examples in natural language. Some of these examples, however, may be inappropriate in

English version, when using pronouns "his/her/ their/its".

First, we notice the lack of gender or number agreement between PPAs and their antecedents. The English version of example 1 has a trivial solution, based on syntactic constraints, but the Portuguese version is ambiguous:

Ex 1: João falou a Maria sobre seu cachorro. (John told Mary about his dog).

Example 2 shows that PPAs can occur in several grammatical (usually, non-subject) positions. Besides, in example 3, we notice that PPAs can refer to different segments of a "NP-of-NP-of-NP..." chain. This kind of structure, with several NPs in the same chain, is typical in our domain.

Ex 2: João viu um cachorro trazendo seu jornal | seu filhote. (John saw a dog bringing his newspaper | its puppy).

Ex 3: O pai do garotinho vendeu sua casa. (The father of the little boy sold his house).
O dono do cão vendeu seu carro | seu filhote. (The owner of the dog sold his car | its puppy).

In some situations, PPAs like shown in example 2 and 3 can be solved by applying semantic knowledge, since PPAs establish possessive relationships (in concrete or figurative sense) between objects in discourse. For example, a human being can usually possess "his car", but a dog cannot. However, we have found in our corpus several PPAs, namely abstract anaphors, which cannot be particularly related to any semantic object. For example, we have PPAs such as "their importance", "their relevance", etc. Similarly, we have found also some abstract antecedents, such as "the problem", "the importance", etc.

Finally, we notice that, in our corpus, we have to treat long and complex sentence structures, which are typical in the domain (laws) that we are dealing with. Thus, despite PPAs in our corpus are mostly (99%) intrasentential, there is a high number of candidates for each anaphor.

3. Factors in PPA resolution

This section describes a minimal set of factors in PPA resolution, based on corpus investigation. These factors will be considered in place of traditional syntactic constraints, which are not suitable for our present problem, as shown in section 2. In our proposal, because of the structural complexity of sentences in the domain, we have adopted a practical approach, based on simple heuristic rules, with a view to avoiding syntactic and semantic analysis. Similar strategies have been adopted in several recent works in anaphor resolution, such as T. Nasukawa (1994), R. Mitkov (1996), R. Vieira & M. Poesio (1997) and others.

We have defined 6 simple factors in PPA resolution (F1 to F6) based on syntactic, semantic and pragmatic knowledge, aiming to determine PPAs antecedents in our specific domain. As a secondary goal, we apply our proposal also to PPAs in a different domain (see section 5). Factors, enunciated as heuristic rules, will act as constraints (F1 to F5) or preferences (F6), as established by J. Carbonell (1988).

3.1. Syntactic level

Since typical syntactic constraints are not suitable for PPA resolution, in our approach we have limited the role of syntactic knowledge to simple heuristic rules based on surface patterns. Surface patterns are typical expressions in the domain, which gave information about PPAs antecedents. To each relevant surface pattern, we have associated a heuristic rule. Some of these

rules can directly elect, with high rate of success, the most probable antecedent, whereas others can only exclude a specific candidate:

F1 - in the pattern <NP and | or PPA>, <NP> must be elected the most probable antecedent of <PPA>. Ex: "John and his dog";

F2 - in the pattern <of NP...of PPA>, <NP> must be elected the most probable antecedent of <PPA>. This rule deals with some cases of syntactic parallelism. Ex: "the death of Suzy, of her children and...";

F3 - in the patter <NP of PPA>, <NP> is not a valid candidate for <PPA>. Ex: in "the death of his son", "death" is not a valid candidate;

F4 - in the pattern <NP of NP of NP... of NP>, only the full chain and the last NP can be considered candidates for PPAs antecedents, i.e., NPs in the middle of the chain can be discarded. This rule adapts the study developed by L. Kister (1995) for NP chains in French, and it constitutes an important mechanism for reducing the high number of candidates in our current problem.

3.2. Semantic level

Heuristic rules based on surface patterns are not sufficient to discriminate among a large set of candidates, as we found in our domain. Thus, we also use semantic knowledge in order to increase the results. Our semantic approach considers possessive relationship rules in the form <Obj1 owns Obj2 >, used to represent "part-of" relationships between typical entities of the domain, according to J. Pustejovsky's (1995) semantic theory. For example, in our corpus some PPAs can be solved with knowledge such as <city owns habitants>, <ecosystem owns natural resources> etc.

In order to apply this kind of knowledge to the whole corpus, we have defined object classes and possible possessive relationships among them. For example, for the anaphor "their hunt" in our corpus, there is a semantic rule which expects only a member of the class <animals> as a suitable antecedent. Typical members of this class would be "birds", "mammals" and all related expressions found in our corpus. Based on this organization we have defined another factor in PPA resolution:

F5 - There must be a valid possessive relationship between a PPA and its antecedent.

3.3. Pragmatic level

Working together, surface patterns and possessive relationships can deal with many PPAs found in our corpus, but we still have two problems to be solved: semantic ambiguity among two or more acceptable candidates and abstract anaphors/antecedents, which cannot be solved by simply applying possessive relationship rules.

For these cases, and possibly for some other cases not included in previous rules, we suggest a pragmatic factor, adapted from S. Brennan's et al (1987) centering algorithm. Although sentence center plays a crucial role in many works in anaphor resolution, usually limiting the number of candidates to be considered, we notice that, because PPAs can refer to almost any NP in the sentence (rather than, for example, personal pronouns, which are often related to the sentence center), pragmatic knowledge plays only a secondary - but still important - role in our approach. We have adapted basic aspects of center algorithm, considering subject/object preference, and domain concepts preference,

suggested by R. Mitkov (1996), aiming to estimate the most probable center for intrasentential PPAs. Thus, in case of ambiguity among candidates (after applying factors F1 to F5), we will consider the estimated center as the preferable PPA antecedent. This constitutes our final rule:

F6 - the sentence center will be preferred among remaining candidates.

4. A distributed architecture for PPA resolution

Factors have been grouped in three knowledge sources (KSs), as part of a blackboard architecture, based on D. Corkill's (1991) work, as shown in figure 1. KSs are independent modules specialized in different aspects of PPA resolution problem (surface patterns, possessive relationships, sentence center), providing both knowledge and procedure distribution among autonomous entities (specialists).

Since in our proposal knowledge and procedure are represented by heuristic rules, KSs have been implemented as reflexive agents, according to S. J Russel & P. Norvig (1995) work. A reflexive agent is a rule-based entity, which acts according to the perceived environment (the blackboard structure).

The blackboard is a global database containing information about the problem (PPA) to be solved: sentence structure information and a set of hypotheses (candidates) to be evaluated by specialists (KSs). The specialists watch the blackboard, looking for a PPA problem to be solved, and evaluate the given data. Specialists can elect, discard or assign preferable candidates, according to their condition-action rules.

The resolution process is coordinated by PPA solver agent, a specialist in PPA resolution. When the PPA solver agent receives a PPA resolution requirement, it writes the initial data (in our current implementation, for intrasentential PPAs, all previous NPs in the sentence are considered as part of the initial set of candidates) onto the blackboard and activates the specialists. After each contribution, the PPA solver evaluates the number of remaining candidates and the possible need for further contributions. At the end of the cycle, in case of ambiguity, the PPA solver will choose the preferred candidate, as determined by the sentence center specialist.

The motivations for adopting a blackboard architecture are the benefits of heterogeneous knowledge distribution and independence among KSs. These benefits will allow us to expand the architecture, adding new factors in PPA resolution or even adding new specialists, dedicated to different anaphoric phenomena.

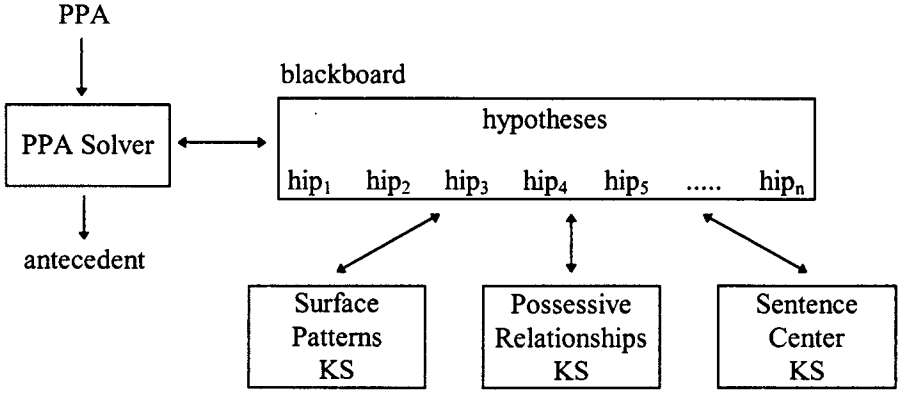

Figure 1 - a distributed architecture for PPA resolution

5. Results

We have examined 198 PPAs from a corpus on Brazilian laws about environment protection. As a result of our implementation, we have achieved a success rate of 92,97%. We evaluate this result as very successful, considering the small set of factors taken in account.

We have also examined PPAs in a second text genre, taking sentences from magazine scientific articles. Within these texts, we have taken 100 intrasentential PPAs, and our strategy has chosen a correct antecedent in 88% of the cases. This deterioration, as a consequence of some different surface pattern occurrences, is to be expected in a new text genre.

As a future work, we aim to expand the architecture, by means of adding new specialists and improving the control mechanism, in order to solve intersentential PPAs and different kinds of pronouns, such as demonstrative and personal.

Acknowledgements

This work was sponsored by CNPq/Protem grant 680081/95-0, as part of the NALAMAS - Natural Language Multi-agent systems - project.

References

Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987) A centering approach to pronouns. In: *proceedings of the 25th ACL.*

Carbonell, J.G. & Brown, R.D. (1988) Anaphora Resolution: A Multi-Strategy Approach. In: *proceedings of COLING'88*, Budapest, Hungary.

Corkill, D.D. (1991). Blackboard Systems. *AI Expert* 6(9):40-47 Sept 91.

Kennedy, C. & Boguraev, B. (1996) Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In: *proceedings of COLING'96*, Copenhagen, Denmark.

Kister, L. (1995) Accessibilité Pronominale des *DÉT. N1 de (DÉT.) N2*: le Rôle de la Détermination. *Lingvisticae Investigationes* XIX:1. 107-121. John Benjamins Publ. Co., Amsterdam.

Mitkov, R. (1996) Attacking Anaphora on All Fronts. In: A. Ramsey (ed.). *Artificial Intelligence: Methodology, Systems, Applications*. IOS Press..

Nasukawa, T. (1994) Robust Method of Pronoun Resolution using Full-text Information. In: *proceedings of COLING'94*, Kyoto, Japan.

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

Russel, S. J. & Norvig, P. (1995) *Artificial Intelligence: a modern approach*. Prentice-Hall, New Jersey.

Sauvage-Wintergerst, C. (1992) *Parallélisme et Traitement Automatique des Langues: Application à l'Analyse des Énoncés Elliptiques*. Université de Paris-Sud (Phd thesis).

Stuckardt, R. (1996) Anaphor Resolution and the Scope of Syntactic Constraints. In: *proceedings of COLING'96*, Copenhagen, Denmark.

Vieira, R. & Poesio, M. (1997) *Processing definite descriptions in corpora*. University of Endinburgh, Scotland.