Anaphora Resolution in a Machine Translation System*

Horacio Saggion[1] & Ariadne Carvalho[2]

[1]Universidad de Buenos Aires, Argentina
hsaggion@dc.uba.ar

[2]Universidade Estadual de Campinas, Brazil
ariadne@dcc.unicamp.br

Abstract

This work is concerned with the automatic translation, based on the transfer approaach, of scientific abstracts in Portuguese.The emphasis is on the resolution of anaphoric references, which are the "bottleneck" of most systems. The analysis component transforms the abstract into an interlingua representation, capturing the essential syntactic and semantic information necessary for translation and resolving the anaphoric references present in the source text. The interlingua representation should be as independent as possible from the language which the abstract will be translated into, since this kind of information is provided by another component. The reason for choosing abstracts is the fact that, although relatively short, they show all the interesting aspects of coherence and cohesion present in larger texts.

1 Introduction

Natural Language is an efficient means of communication because of the assumptions that the speaker and the hearer can make about each other [CD87]. The hearer assumes that what the speaker is trying to communicate forms a coherent whole, rather than a set of isolated and unrelated comments; and the speaker assumes that the hearer is making an assumption and is, therefore, trying to interpret what he/she hears as a coherent message. This allows the speaker to make use of abbreviated linguistic forms, expecting that the hearer will recognise and interpret them with reference to what was previously said. Anaphora is a special case of cohesion. It can be defined as the device of making an abbreviated reference to some entity (or entities) with the expectation that the perceiver of the discourse be able to disabbreviate the reference and, thereby, determine the identity of the entity. The abbreviated reference is called an anaphor and the entity to which it refers is its referent or antecedent [HG81]. The process of determining the referent of an anaphor is called resolution.

The work described here is concerned with the translation of abstracts from scientific papers. The task can be divided into two main steps: analysis and synthesis. We concentrate on the

*This research was partly sponsored by grants from the Brazilian financial agencies CNPq and FAPESP.

first. Specifically, we are interested in the resolution of anaphoric references, which are the "bottleneck" of most natural language processing systems.

Hirst describes twelve types of anaphora. We will be dealing with one of them, namely *definite pronominal reference*. The word pronoun has two meanings. Firstly, it can refer to a part of speech such as "he", "she", "it", "they" or "that". Secondly, it can refer to an anaphor whose antecedent is a noun phrase, that is one which "stands in place of a noun". In this work we will be dealing with both.

The remainder of the paper is organised as follows. In Section 2 we will define abstracts and their function; in Section 3 we will define reference and describe approaches which have been used to deal with it. In Section 4 we will describe possible architectures for Machine Translation (MT) systems. In Section 5 we will introduce the system proposed; in Section 6 we will discuss possible internal representations to be used by the system and, finally, in section 7 we will present our conclusions.

2 Abstracts and their Use

An abstract is the first section of a report, coming after the title and before the introduction. It provides the reader with a brief preview of the paper. Many readers depend on the abstract to give them enough information about the work in order to decide if they will read the entire report or not. Abstracts from almost all fields are written in a very similar way. The type of information included and the order the information is presented are very conventional. Abstracts are written to be as brief and concise as possible.

According to the Brazilian Technical Norms Association (ABNT) and according to [WR90] an abstract may include the following elements:

- background: where, as the name says, background information must be given:
- purpose: where the objectives of the work must be described;
- method: where new techniques and methodological principles are described;
- results: where new facts are presented;
- conclusion.

Sometimes, for journal articles the editor establishes a word limit for the abstract. When this happens, the authors can combine information, such as the purpose and method of the study, in one sentence.

As well as the norms on which elements to include in the abstract, there are also language conventions. The most common is the convention on verb uses, which are: in the purpose item the verbs should be in the present tense; in the method and result items the verbs should be in the past tense; in the conclusion either the verbs could be in the present tense or modal auxiliaries might be used. Consider the following abstract, taken from the "Revista de Ensino de Engenharia" [EN85]. We have chosen to study abstracts from this magazine because submissions must be in accordance with the norms from ABNT.

Uma experiência no ensino de Cálculo Numérico na UFSC[1]

> Com a utilização de calculadoras programáveis e microcomputadores, torna-se necessária a adequação do plano de ensino de Cálculo Numérico para a formação dos futuros engenheiros. Algumas alterações na metodologia de ensino **desta** disciplina são sugeridas e feitas algumas recomendações quanto à utitização de calculadoras programáveis.

An experience on teaching Numerical Calculus at UFSC

> *With the increasing utilization of programmable calculators and microcomputers, it becomes necessary to provide adequate teaching plans for numerical calculus instruction at the undergraduate level in engineering schools. Some changes in the methodology for teaching* **these** *subjects are suggested and some recommendations are made for the utilization of programmable calculators.*

There are two types of information present in the abstract: *purpose* and *method*. The first sentence contains the purpose of the abstract and the second contains the method used to achieve it. Also, following the norms from ABNT, the verb "torna-se" ("it becomes") in the first sentence is in the present tense and the verbs "são sugeridas" ("are suggested") and "feitas" ("are made") are in the past tense.

3 How to Resolve Reference

We need two different techniques for handling anaphora: one to handle intrasentential uses (that is, when the anaphor and the referent are in the same sentence) and one to handle intersentential uses (that is, when the anaphor and the referent are not in the same sentence).

3.1 Intrasentential Reference

There are other significant syntactic constraints on what objects may be referents for a pronoun. In particular, the following condition must usually hold:

(I) A pronoun and its antecedent must agree in number, gender and person.

There are other syntactic constraints based on the structure of the sentence. Consider the following example:

(a) Maria viu-se no espelho.
Mary saw **herself** *in the mirror.*

The derivation tree for this sentence is shown in Figure 7 - 1 (a).

[1] The examples used throughout the paper will be in Portuguese, each followed by its translation in English.

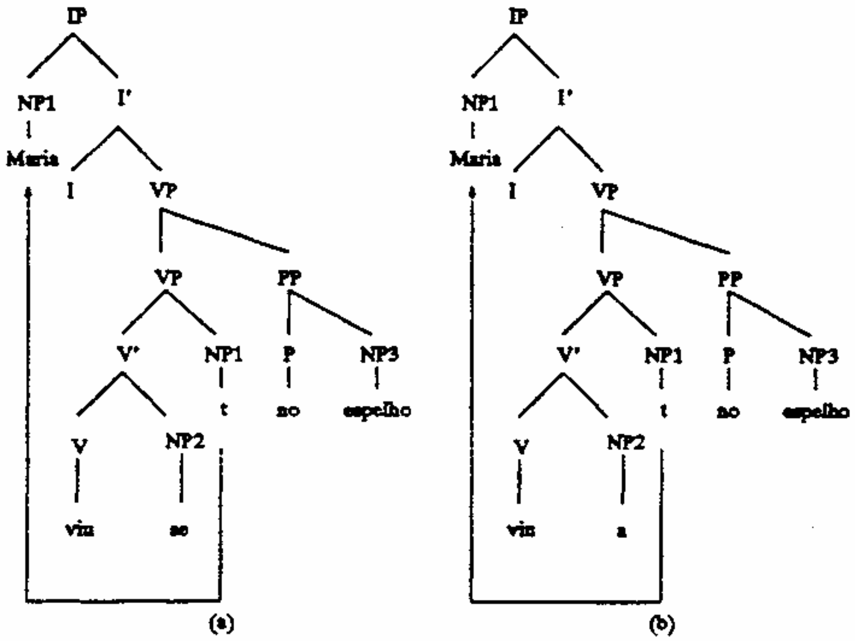

Figure 7 - 1 Derivation Tree

In this case pronoun "se" has as its only possible antecedent "Maria". We have to resolve the reference because in Portuguese the reflexive pronoun "se" is used in the masculine, feminine and neuter cases. If we do not resolve it, when the time for generation comes the generator would not know what the translation for "se" would be: "herself", "himself" or "itself"? The resolution of this case is based on the notion of *constituent-command (c-command)*, which is a relation among the nodes of a tree structure and which is defined as follows [RT83][2]:

> *Node A c-commands node B iff the first branching node α_1 dominating A dominates B.*

The c-command based constraints on coreference are:

(II) A pronoun must be interpreted as non-coreferential with any full noun phrase (NP) that it c-commands.

(III) A reflexive pronoun must be interpreted as co-referential with (and only with) a c-commanding NP within a specific syntactic domain (e.g. its minimal governing category)[3].

(IV) A non-reflexive pronoun must be interpreted as non-coreferential with any c-commanding NP in the syntactic domain which is specified for (III).

[2] Alternative definitions of c-command can be encountered in the literature.

[3] The minimal governing category (MGC) of a node α is the minimal (lowest) S or NP dominating α.

Therefore, in sentence (a) pronoun "se" is c-commanded by the NP "Maria" and, according to constraint (III), it must be the referent of the pronoun. On the other hand, in the sentence:

(b) Maria viu-**a** no espelho.
Mary saw **her** *in the mirror.*

constraint (IV) forbids pronoun "a" to refer to "Maria", because it is being c-commanded by the NP. In this case. the referent must be found elsewhere in the text. The derivation tree for this sentence is shown in Figure 7 - 1 (b).

3.2 Intersentential Reference

The main technique for handling intersentential anaphora is the maintenance of a record of all objects mentioned in the preceding sentences. This record consists of the syntactic analysis of the immediately preceding sentence, plus an ordered list of all referents mentioned in the last several sentences, called *history list* [AJ87].

So, for example in the case of sentence (b), we would search the history list, starting from the most recently mentioned objects, until we find one that satisfies the number, person and gender information of the pronoun.

Nonetheless, sometimes syntactic constraints alone are not able to resolve the reference. Consider the following sentence:

(c) João não deu os pêssegos aos pássaros. Embora **eles** estivessem famintos, **eles** estavam verdes.
John didn't give the peaches to the birds. Although **they** *were hungry,* **they** *were ripe.*

If we only take syntactic constraints into account, both "pêssegos" and "pássaros" could be the referents for both "eles". But if we interpret them in the light of semantic preferences, using the knowledge that "pássaros", being animate, are likely to be hungry and "pêssegos", being fruits, are likely to be ripe, we encounter the right referents.

But there is also another problem. Consider again the abstract shown in Section 2. Note that the definite reference "desta disciplina" ("this discipline") needs an antecedent. The problem is that the antecedent is "Cálculo Numérico" (" Numerical Calculus") which, in Portuguese, is a masculine noun, as opposed to "desta disciplina", which is a feminine noun. Therefore, constraints on gender and number would not help to find the referent. What we really need here is an inference mechanism, through which it could be inferred that "Cálculo Numérico" is a discipline and, as such, can be the referent of "desta disciplina".

4 Possible Architectures for Machine Translation Systems

Direct Systems use neither syntactic nor semantic information. They are generally designed for a specific pair of languages: the source language (SL) and the target language (TL). They use pattern matching to identify and transform parts of the speech and to produce the translation from the original text. For example, in Portuguese it is possible to have a noun phrase composed of a noun followed by several adjectives, as can be seen next:

(c) A garota bonita e inteligente.
The beautiful and intelligent girl.

In English, the adjectives which qualify the noun must precede it. Direct Systems use rules to transform the Portuguese pattern into the English one. The translation generated by these systems usually requires a revision before the user is presented with the final version [TA84].

In the *Transfer Approach* text processing is done in tree stages: analysis, transfer and generation. In the analysis stage the original text is processed and an internal representation (close to the SL) is produced, as a result of syntactic and semantic analysis; here nothing is known about the target language. In the generation stage a translation, based on an internal representation (close to the TL) and on a set of generation rules, is produced; here nothing is known about the source language. The transfer stage is responsible for bridging the gap between the two internal representations.

AI-Based Systems rely on a "complete" understanding of the text before translation takes place. They rely on an *interlingua* representation, which is independent of any natural language, to represent the meaning of the text. The processing is characterized by the lack of syntactic rules and generally it is based on semantic information. Many systems using this paradigm were constructed, but in general the text processed by them is restricted to very stereotypical subjects [SR84].

5 A Machine Translation System

Direct systems have no flexibility when an extension is needed in order to cope with more than one pair of languages. In general, they do not resolve pronominal references, relying on the revisor for that. Systems that completely "understand" the text before synthesis begins are interesting but they need very large knowledge bases to be able to process the input. The system proposed here is intended to be hybrid, as it will make use of the transfer approach with the incorporation of some Artificial Intelligence (AI) techniques as a way of obtaining a high quality translation with minor effort. The complete understanding of a text is a complex task; in our system, understanding means identifying and resolving all referents present in the original text.

5.1 The Architecture of the System

The system's architecture can be seen in Figure 7 - 2. Text processing is accomplished in two stages: analysis and synthesis.

The *analysis* step is responsible for the generation of a structure which represents the meaning of the abstract. This representation must contain the actions and states described, which in the text are realized as verbs, and all the actors involved, which in the text are realized as noun phrases.
This step is also broken into morphological and syntactical analysis and semantic interpretation. Portuguese morphology is very rich, specially with respect to verbs. Consider, for example, the indicative mood for the verb "amar" (to love), which is a regular verb:

- Present (indicative mood)

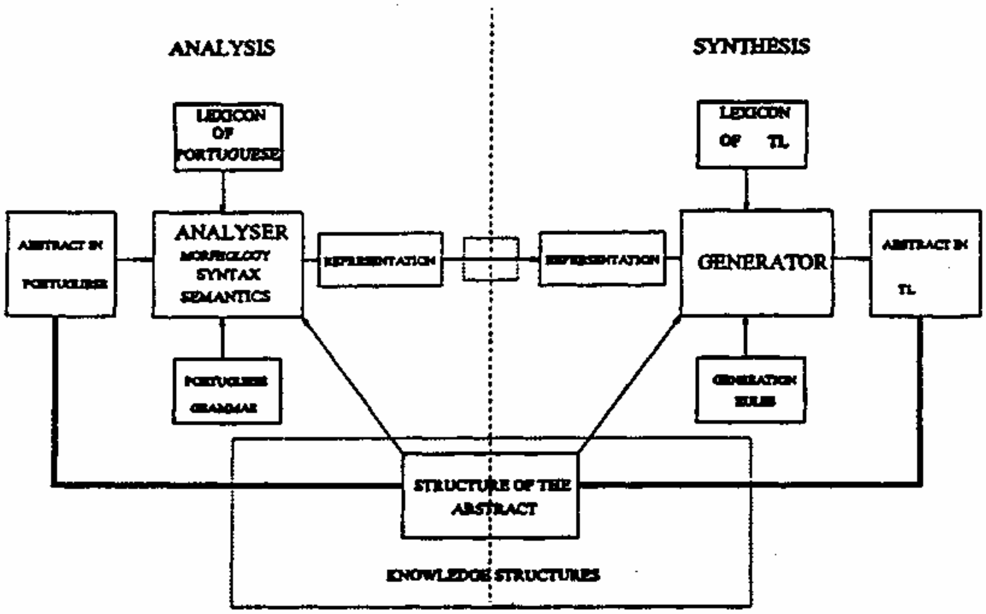

Figure 7 - 2 A Machine Translation System

- eu **amo** (I love)
- tu **amas** (you love)
- você **ama** (you love)
- ele/ela **ama** (he/she loves)
- nós **amamos** (we love)
- vós **amais** (you love)
- vocês **amam** (you love)
- eles/elas **amam** (they love)

- Imperfect (indicative mood)
  - eu **amava** (I loved)
  - tu **amavas** (you loved)
  - você **amava** (you loved)
  - ele/ela **amava** (he/she loved)
  - nós **amávamos** (we loved)
  - vós **amáveis** (you loved)
  - vocês **amavam** (you loved)
  - eles/elas **amavam** (they loved)

As can be seen, in Portuguese every person in the "presente" tense has a different form, as opposed to English, where there are only two possible forms. In the "imperfeito" tense, there are five different forms, as opposed to English, where there exists only one.

To deal with such complexity a *morpho-lexical analyser* will be used, to break the text into tokens, based on morphological rules as well as on the Portuguese dictionary. These tokens, containing syntactic and semantic information, will be the input of the syntactic analyser.

The *syntactic analyser* is responsible for the creation of a structural representation for each sentence in the form of a tree which relates parts of the speech. Our approach to syntactic analysis is based on logic grammars because it allows the automatic construction of the tree as interpretation proceeds [MM89].

The *semantic interpreter* is responsible for the right association between the actions and the actors involved; it is responsible for the assignment of correct roles to the entities in the sentence. The syntactic analyser and the semantic interpreter will interact in order to reject an incorrect analysis of the sentence as soon as possible.

Consider the following examples:

(d) João deu o livro de Maria para Pedro.
John gave Mary's book to Peter.

(e) João deu o livro de texto para Pedro.
John gave the text book to Peter.

In the first case "Maria" is the owner of the book and in the second, "texto" is an attribute of book. A syntactic analyser alone will not be able to differentiate between the prepositional phrases "de Maria" and "de texto"; therefore, an interaction between the syntactic and semantic components is necessary, in order to allow the correct translation of sentences (d) and (e) into English. The semantic interpreter must also resolve the references which have not been resolved by the syntactic analyser. The antecedent of an anaphor will be one of the entities saved by the syntactic analyser and the choice will be based not only on syntactic agreement, but also on semantic constraints.

As the two internal representations are usually different, a *transfer* stage will always be needed to transform the output of the analyser into the input of the synthesis stage. In the case of close languages such as the pair Portuguese-Spanish this stage will be very simple, but as the "distance" between the languages increases this becomes more and more complex.

The *synthesis* step must generate the abstract in the target language according to the target language internal representation produced by the transfer component; in this stage knowledge on the structure of the abstract is used in order to correctly generate the information present in the original text. Generation rules must deal with aspects of discourse, that is, they must take into account the fact that scientific abstracts are being generated and, therefore, they should follow the norms presented in Section 2. Furthermore, they must generate the references in the target language to account for the cohesive aspects of the text.

6 Possible Internal Representations

Not only AI-based systems, but also systems based on the transfer approach, use an internal representation to capture the "meaning" of the sentences. We believe that an acceptable quality of the text may be achieved by means of an internal representation close to the source language (in this case Portuguese). This representation must allow not only for the description of every single sentence from the text, but also it must represent the relationships between them.

Many different meaning representations were used by natural language processing systems. Semantic networks were initially used with this purpose [SR73]. They are very useful because the relationships between the actions and the actors can be well represented within this formalism. They can also be used as input for an Augmented Transition Network (ATN) [BM78] that generates sentences in natural language.

Some systems use D-structure [RA88, IP85] as an internal representation of the sentence, as it carries both syntactic and semantic relations between parts of the sentence. Transformation rules can then be applied to produce an S-structure of the sentence.

First order logic was extensively used in AI-applications to represent the meaning of the sentence through predicates and arguments [PF83]. A dictionary that relates the lexicon to the predicates can be used to generate a sentence in the TL according to the generation rules.

Functional Unification Formalism (FUF) [EM92] uses feature structures to represent the meaning of the sentence and a grammar is used to produce text in English, having the feature structures as its input.

Frames were used in AI-based systems to do multilingual translation [RS90]; in this formalism slots and values are used to represent the meaning of the sentence.

Conceptual Dependency Theory [SR77] makes it possible to construct the structure of the sentence based on primitive actions and actors. Scripts provide a way of understanding texts about stereotypical events like "eating at a restaurant" [SR77]. Systems using this kind of representation were successfully used because they showed some understanding through the translation of the original text.

We have concentrated on the construction of a formalism which must be able to represent not only the relationships between the actions and actors involved in the sentences, but also the coherence relations that exist among the parts of the text.

The representation must be able to correctly capture the pattern of the abstract, according to the norms described in Section 2. This information will help during the target language generation because it will assist in the lexical choices that the generator will have to make.

The internal representation formalisms presented here were mostly used to represent the meaning of isolated sentences; we intend to extend the one which we find to be most appropriate to represent the meaning of the entire text.

7 Conclusions

Analysis and translation are very complex tasks. In order to deal with such complexity we have chosen to work with abstracts from scientific papers because, although being as concise as possible, they must still be written according to the coherence and cohesive relations present in larger texts.

In order to produce high quality texts an understanding of the original text is necessary and there is no understanding without anaphora resolution. Therefore, the system proposed here concentrates on this aspect of text processing.

Although we have already identified the transfer approach as the more appropriate for the work which is being developed, we have not yet defined the internal representation which will be used by the system.

We think that none of the existing representations will be able to deal with the entire text, since they were designed to deal with isolated sentences. Thus, the development of such a formalism will be one of the main goals of this work.

References

[AJ87] Allen, J. *Natural Language Understanding.* The Benjamin Cummings Publishing Company, Inc., 1987.

[BM78] Bates, M. *The Theory and Practice of ATN Grammars.* Natural Language Communication with Computers. Lecture Notes in Computer Science, Number 63. Springer-Verlag 1978.

[CD87] Carter, D. *Interpreting Anaphors in Natural Language Texts.* Ellis Horwood Limited, 1987.

[EN85] de Araújo, N. D. Uma Experiência no Ensino de Cálculo Numérico na UFSC. *Revista de Ensino de Engenharia*, 4(2): 138-139, 1985.

[EM92] Elhadad, M. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation.* Phd. Thesis. Graduate School of Arts and Sciences, Columbia University, 1992.

[HG81] Hirst, G. Anaphora in Natural Language Understanding: A Survey. *Lecture Notes in Computer Science 119.* Springer-Verlag, 1981.

[IP85] Isabelle, P. and L. Bourdeau. *TAUM-AVIATION: Its Technical Features and Some Experimental Results.* Computational Linguistics. Vol. 11, Num 1, January-March 1985.

[MM89] McCord, M. *Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars.* Lecture Notes in Natural Language and Logic, Number 459, Springer-Verlag 1989.

[PF83] Pereira, F.C.N. *Logic for Natural Language Analysis.* SRI International. Technical Note 275, 1983.

[RA88] Radford, A. *Transformational Grammar: A First Course.* Cambridge Textbooks in Linguistics. Cambridge University Press. 1988.

[RS90] Raman, S. and N. Alwar. *An AI-Based Approach to Machine Translation in Indian Languages.* Communications of the ACM. Vol. 33. Num. 5. May 1990.

[RT83] Reinhart, T. *Anaphora and Semantic Interpretation.* Croom Helm Ltd., 1983.

[SR77] Schank, R., R. Abelson. *Scripts Plans Goals and Understanding.* Lawrence Erlbaum Associates, Publishers, 1977.

[SR84] Schank, R., P.G. Childers. *The Cognitive Computer.* Addison-Wesley Publishing Company, Inc., 1984.

[SR73] Simmoms, R.F. *Semantic Networks: Their Computation and Use for Understanding English Sentences.* Computer Models of Thought and Language, R. Schank, K. Colby Eds, 1973.

[TA84] Tucker, A.B. *A Perspective on Machine Translation: Theory and Practice.* Communications of the ACM. Vol. 27. Num 4. April 1984.

[WR90] Weissberg, R. and S. Buker. *Writing up Research.* Prentice-Hall, Inc..