Hobbs' Algorithm for Pronoun Resolution in Portuguese

Denis Neves de Arruda Santos and Ariadne Maria Brito Rizzoni Carvalho

Institute of Computing, State University of Campinas,
Caixa Postal 6176, 13083-970, Campinas, SP, Brazil
denis.santos@students.ic.unicamp.br,ariadne@ic.unicamp.br

Abstract. Automatic pronoun resolution may improve the performance of natural language systems, such as translators, generators and summarizers. Difficulties may arise when there is more than one potential candidate for a referent. There has been little research on pronoun resolution for Portuguese, if compared to other languages, such as English. This paper describes a variant of Hobbs' syntactic algorithm for pronoun resolution in Portuguese. The system was evaluated comparing the results with the ones obtained with another syntactic algorithm for pronoun resolution handling, the Lappin and Leass' algorithm. The same Portuguese corpora were used and significant improvement was verified with Hobbs' algorithm.

1 Introduction

Anaphora is a linguistic phenomenon of making an abbreviated reference to some entity (or entities) expecting the perceiver of the discourse to be able to disabbreviate the reference and, therefore, 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. The process of determining the referent of an anaphor is called resolution [8]. One of the most common types of anaphora is pronominal anaphora. Difficulties may arise when there is more than one potential candidate for a referent. Consider the following sentence:

(1) *João culpou Pedro por ter batido seu carro.*
[John blamed Peter for crashing his car.]

The anaphor "*seu*" (his) may either refer to "*João*" (John) or to "*Pedro*" (Peter).

Automatic pronoun resolution may improve the performance of natural language systems, such as translators, generators and summarizers. Several algorithms have been proposed to deal with the problem, such as Hobbs' algorithm, [9], Centering algorithm [7], and Lappin and Leass' algorithm [10]. Hobbs developed two approaches to pronoun resolution: a syntactic and a semantic approach.

In this paper we describe a variant of Hobbs' syntactic algorithm for pronoun resolution in Portuguese. Hobbs' algorithm was chosen due to its simplicity and good performance with English texts. The system was evaluated comparing the

A. Gelbukh and A.F. Kuri Morales (Eds.): MICAI 2007, LNAI 4827, pp. 966–974, 2007.

results with the ones obtained with Lappin and Leass' algorithm. The same Portuguese corpora were used with both algorithms, and better performance was verified with Hobbs' algorithm.

The remainder of the paper is organized as follows: in Sect.2, an overview on pronoun resolution is given; in Sect.3, a variant of Hobbs' algorithm, followed working examples, is presented; in Sect.4, Hobb's algorithm is evaluated on three different corpora, and a comparison with Lappin and Leass' algorithm is made; finally, in Sect.5, the conclusions and future work are presented.

2 Related Work

There have been a few approaches to pronoun resolution in Portuguese. In [14], the authors have proposed an algorithm for possessive pronoun handling in Portuguese. Their strategy was to use syntactic, semantic and pragmatic knowledge. The algorithm was evaluated using the text of Brazilian Laws on Environment Protection and, according to the authors, it succeeded in 92.97% of the cases.

In [1], the authors have evaluated the Centering algorithm for pronoun resolution in Portuguese texts. Centering algorithm [7] is based on a system of rules and restrictions that govern the relations between referring expressions. The authors used 16 juridical texts from the Attorney General's Office of the Republic of Portugal to evaluate the system and, according to them, the algorithm succeeded in 51% of the cases.

In [4,5] a variant of Lappin and Leass' algorithm for pronoun resolution in Portuguese was proposed. Lappin and Leass' algorithm [10] deals with third person pronouns in English; it uses a syntactic representation generated by a parser, and salience measures derived from the syntactic structure of the sentence. The algorithm was evaluated on three different corpora: 14 texts from a Brazilian magazine; a well-known literary book called "O Alienista"[2]; and the same 16 juridical texts used to evaluate the Centering algorithm [1]. For the magazine texts, the algorithm succeeded in 43.56% of the cases; for the literary book, the success rate was 32.61%; and, for the juridical texts, the algorithm succeeded in 35.15% of the cases.

In [13] the authors have proposed an algorithm for identifying noun phrase antecedents of third person personal pronouns, demonstrative pronouns, reflexive pronouns, and omitted pronouns in unrestricted Spanish texts. The authors defined a list of constraints and preferences for different types of pronominal expressions. The algorithm was evaluated on a corpus of 1,677 pronouns and achieved a success rate of 76.8%. The authors implemented other four competitive algorithms, including Hobbs' algorithm, and tested their performance on the same test corpus. The proposed algorithm obtained the best results.

3 Hobbs' Syntactic Algorithm

The syntactic algorithm developed by Hobbs solves intra- and inter-sentencial pronominal anaphora. The strategy is a special traversal of the surface parse tree[1], looking for a noun phrase of the same gender and number of the pronoun [9]. The algorithm also deals with cataphora; however, it does not handle reflexive pronouns, and sentences which are themselves antecedents.

The tree is traversed in a breadth-first, left-to-right manner, because it is more likely that pronouns refer to subjects, instead of objects.

For convenience, Hobbs [9] makes the assumption that pronouns are immediately dominated by an NP node; hence, possessive nouns and pronouns have the structure illustrated in Fig.1[2].

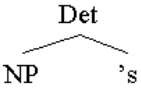


Fig. 1. Structure of a possessive noun or pronoun

Other assumption made by Hobbs is that an N within an NP node may have a prepositional phrase attached to it. This assumption is made to distinguish between the following two sentences:

(2) Mr. Smith saw a driver in his truck.
(3) Mr. Smith saw a driver of his truck.

In (2), pronoun "his" may either refer to the driver or to Mr. Smith; however, in (3) it can only refer to Mr. Smith. Figures 2(a) and 2(b) illustrate the corresponding parse trees for (2) and (3), respectively.

3.1 The Algorithm

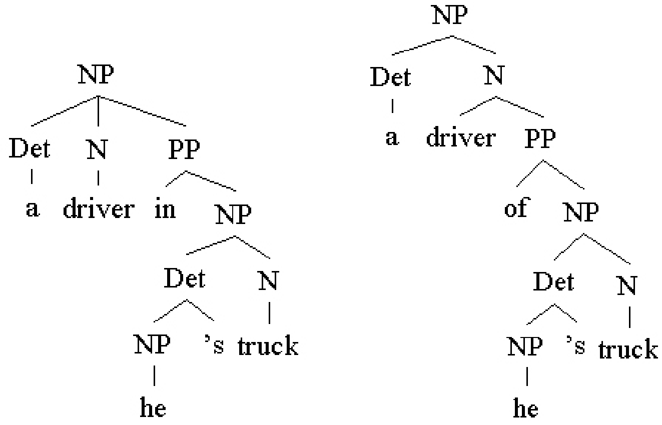
Hobbs' algorithm is based on a particular traversal of the parse tree, looking for a noun phrase of the correct gender and number. Some constraints on anaphora resolution are encoded into the control structure of the algorithm [15]. The original algorithm was extended to deal with reflexive pronouns.

The algorithm takes as input a pronoun and a parse tree, and returns an antecedent to the pronoun. We will say that a node is *acceptable* if it agrees in gender and number with the pronoun. The variant of the algorithm is stated as follows, with the convention that all tree traversals are in left-to-right, breadth-first order:

1. Starting at the *NP* node which immediately dominates the pronoun, go up the tree to the first *NP* or *S* node encountered. Call this node *X*, and call the path used to reach it *p*.
2. (Reflexive case) If the pronoun is reflexive, traverse all branches below node *X* to the left of path *p*, and return the first acceptable *NP* node found.

[1] The tree that exhibits the grammatical structure of the sentence.

[2] Adv: adverb; Art: article; Conj: conjunction; Det: determiner; N: noun; NP: noun phrase; PP: prepositional phrase; Prp: preposition; S: sentence; VP: verb phrase.



(a) Surface parse tree of "*a driver in his truck*".

(b) Surface parse tree of "*a driver of his truck*".

Fig. 2. The corresponding parse trees for (2) and (3)

3. Traverse all branches below node X to the left of path p and return the first acceptable *NP* node encountered which has an *NP* or S node between it and X.
4. While X is not the highest S node in the sentence, repeat:
   (a) From node X, go up the tree to the first *NP* or S node encountered. Call this new node X, and call the path traversed to reach it p.
   (b) If X is an acceptable *NP* node, and if the path p to X did not pass through the N node that is immediately dominated by X, return X.
   (c) Traverse all branches below node X to the left of path p, and return the first acceptable *NP* node found.
   (d) If X is an S node, traverse all branches of node X to the right of path p, but do not go below any *NP* or S node encountered. Return the first acceptable *NP* node found.
5. Traverse the surface parse tree for the previous sentences in reverse order of occurrence proposing the first acceptable *NP* node found.

Reflexive pronouns are dealt with in step 2; step 5 is concerned with intersentencial anaphora, that is, antecedents which belong to previous sentences. Hobbs [9] suggests to restrict the search to a five-sentence window. In step 4.b there is an embedded constraint that takes care of the case shown in Fig.2(b). Step 4.d handles the cases where the antecedent commands but does not precede the pronoun, that is, cataphora [9].

3.2 An Example

We follow the steps of the algorithm on three examples. Consider the following sentence, whose parse tree is shown in Fig.3.

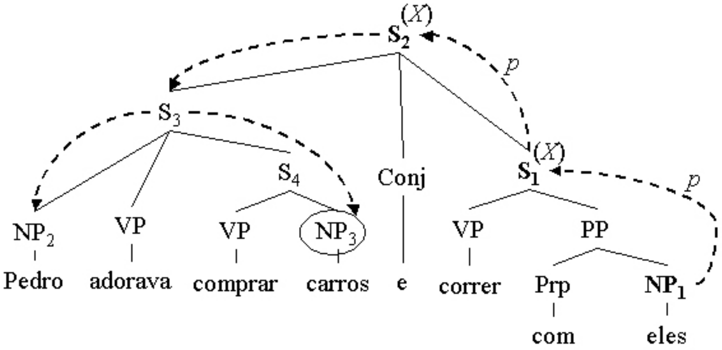


Fig. 3. Illustration of the algorithm

(4) *Pedro adorava comprar carros e correr com eles.*
[Peter loved buying cars and racing them.]

According to step 1, execution begins at node NP$_1$; node S$_1$ is visited and becomes node *X*. Since the pronoun is not reflexive, according to step 3, the left portion of S$_1$ tree is traversed, but no NP node is found. The loop in step 4 is entered, since S$_1$ is not the highest S node in the sentence. According to step 4.a, node S$_2$ is visited, and becomes node *X*. Step 4.b does not apply because node *X* (S$_2$) is not an *NP*. According to step 4.c, the left portion of (S$_2$) is traversed; and node NP$_2$ is the first *NP* found, but it is not acceptable (due to number). The next *NP* node in the traversal is NP$_3$, which is returned.
Now consider (5), which contains a reflexive pronoun "se"(himself):

(5) *O goleiro se machucou.*
[The goalkeeper hurt himself.]

The parse tree for (5) is shown in Fig.4.

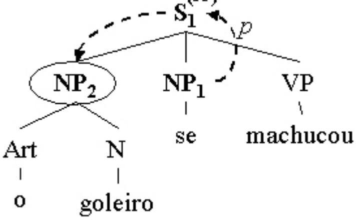


Fig. 4. Illustration of the algorithm with a reflexive pronoun

According to step 1, execution begins at node NP$_1$; node S$_1$ is visited, and becomes node *X*. According to step 2, the left portion of S$_1$'s tree is traversed, and node NP$_2$ is returned as antecedent.

Finally, consider (6), whose parse tree is shown in Fig.5. There are two possible candidates for referent, "João" and "Pedro".

(6) *João deu uma maçã para Pedro. Ele também deu uma melancia.*
[John gave an apple to Peter. He also gave an watermelon.]

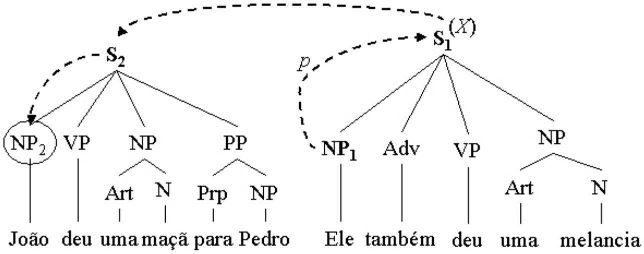


Fig. 5. Illustration of the algorithm with two possible referents

According to step 1, execution begins at node NP$_1$; node S$_1$ is visited, and becomes node *X*. According to step 3, the left portion of the S$_1$ tree is traversed, but no node NP is found. Since S$_1$ is the highest node in the tree, the previous sentence is traversed, according to step 5, and node NP$_2$ is returned as antecedent.

4 Results

The algorithm was tested and evaluated on three different corpora. The first corpus was composed of 14 texts from a Brazilian magazine; the second was a literary book called *O Alienista*, from a well-known Brazilian author, Machado de Assis [2]; and the third was composed of 16 legal opinions from the Attorney General's Office of the Republic of Portugal. The first corpus contained 196 pronouns, 92 being reflexive pronouns; the second corpus contained 634 pronouns, 119 being reflexive pronouns; and, finally, the last corpus contained 297 pronoun anaphora, with no reflexive pronouns.

The same corpora were used to evaluate the variant of Lappin and Leass' algorithm and Hobbs'algorithm. The parser PALAVRAS [3], a robust parser for Portuguese texts, was used to automatically annotate the corpora with morphosyntactic information. Pronouns were manually annotated with a tool for discourse annotation, MMAX (*Multi-Modal Annotation in Xml*) [12]. Another tool, Xtractor [6], was used to generate the XML encoding for the PALAVRAS output to improve the linguistic information extraction from the corpora analysed by PALAVRAS. The tool produces three XML files: *words*, which contains a list of the words from the text; *pos*, which contains morphosyntactic information; and *chunks*, which contains information on the text structure.

We have made some adjustment in the files generated by Xtractor and MMAX to make them suitable for the evaluation process.

In the experiments with the magazine and the literary corpus, we considered the resolution process successful if the solution offered by the algorithm was the same as that offered by manual annotation, or if there was a noun phrase which coreferred with the solution given by the annotators. Therefore, all the algorithm solutions were manually compared with the solution given by the annotators. With the juridical corpus, we considered the resolution process successful if the solution offered by our algorithm was the same as that offered by manual annotation. Therefore, all solutions given by the annotators were later automatically compared with the solutions given by our system.

Table 1 presents the results for the magazine texts; Table 2 presents the results for the literary book; and Table 3 presents the results for the juridical texts. The first column of the table shows the pronoun category (reflexive, non-reflexive and the total); the second shows the success rate for Hobbs' algorithm; and the third shows the success rate for Lappin and Leass' algorithm.

The results have shown that Hobbs' syntactic algorithm for pronoun resolution worked surprisingly well for Portuguese, despite its simplicity. This becomes more evident when compared with Lappin and Leass' algorithm, which is more elaborated. Our variant of Hobbs' algorithm performed significantly better than Lappin and Leass', except for a curious coincidence — the success rate of both algorithms for non-reflexive pronouns on the Brazilian magazine corpus was exactly the same.

Note the remarkable success achieved with the simple device of treating reflexive pronouns separately.

Compared with the results reported by Hobbs on English texts (88.3% of success rate), the numbers do not seem so impressive. Notice, however, that our

Table 1. Results on Brazilian magazine corpus

Pronoun	Quantity	Hobbs' algorithm	Lappin and Leass' algorithm
Reflexive	92	66 (71.74%)	35 (38.04%)
Non-reflexive	104	53 (50.96%)	53 (50.96%)
Total	196	119 (61.22%)	88 (44.90%)

Table 2. Results on literary corpus

Pronoun	Quantity	Hobbs' algorithm	Lappin and Leass' algorithm
Reflexive	119	82 (68.91%)	41 (34.45%)
Non-reflexive	515	233 (45.24%)	174 (33.79%)
Total	634	315 (49.68%)	215 (33.91%)

Table 3. Results on juridical corpus

Pronoun	Quantity	Hobbs' algorithm	Lappin and Leass' algorithm
Non-reflexive	297	120 (40.40%)	103 (35.15%)

tests were more extensive, on a substantially larger corpus. The complexity of available sentence forms in Portuguese should also be taken into account as a possible source of varying performance of the algorithm.

5 Conclusion and Future Work

We have developed a variant of Hobbs' algorithm for pronoun resolution in Portuguese. It was evaluated on three large corpora against Lapppin and Leass' algorithm, and showed a markedly better performance.

As future work, we will introduce semantic knowledge through the application of selectional constraints. Hobbs reported an increase of accuracy to 91.7% with the use of semantics, and we believe a similar improvement will be obtained for Portuguese.

Acknowledgments. We thank CNPq for financial support. We are grateful to Thiago Thomes Coelho and Arnaldo Mandel for their valuable help.

References

1. Aires, A.M., Coelho, J.C.B., Collovini, S., Quaresma, P., Vieira, R.: Avaliação de Centering em Resolução Pronominal da Língua Portuguesa. Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués. Iberamia (2004)
2. Assis, M.: O Alienista. VirtualBooks Literatura Brasileira. VirtualBooks (2002), `http://virtualbooks.terra.com.br/freebook/port/O_Alienista.htm`
3. Bick, E.: The parsing system PALAVRAS: Automatic grammatical analysis of Portuguese in a constraint grammar framework. Årthus University (2000)
4. Coelho, T.T.: Resolução de anáfora pronominal utilizando o algoritmo de Lappin e Leass. Master's thesis. University of Campinas (2005)
5. Coelho, T.T., Carvalho, A.M.B.R.: Lappin and Leass' Algorithm for Pronoun Resolution in Portuguese. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 680–692. Springer, Heidelberg (2005)
6. Gasperin, C.V., Vieira, R., Goulart, R.R.V., Quaresma, P.: Extracting XML chunks from Portuguese corpora. In: Proceedings of the Workshop on Traitement Automatique des Langues Minoritaires. Batz-sur-Mer (2003)
7. Grosz, B.J., Weinstein, S., Joshi, A.K.: Centering: A Framework for modeling the local coherence of discourse. In: Computational Linguistics, vol. 21, pp. 203–225. MIT Press, Cambridge (1995)
8. Hirst, G.: Anaphora in Natural Language Understanding. LNCS, vol. 119. Springer, Heidelberg (1981)
9. Hobbs, J.R.: Pronoun Resolution. Technical Report. City University of New York (1976)
10. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. In: Computational Linguistics, vol. 20, pp. 535–561. MIT Press, Cambridge (1994)
11. Mitkov, R.: Anaphora Resolution. Longman (2002)
12. Müller, C., Strube, M.: MMAX: A tool for the annotation of multi-modal corpora. In: The 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems IJCAI, pp. 45–50 (2001)

13. Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., Muñoz, R.: An algorithm for anaphora resolution in spanish texts. In: Computational Linguistics, vol. 27, pp. 545–567. MIT Press, Cambridge (2001)
14. Paraboni, I., Lima, V.L.S.: Possessive pronominal anaphor resolution in Portuguese written texts. In: The 17th International Conference on Computational Linguistics. Computational Linguistics, vol. 2, pp. 1010–1014. MIT Press, Cambridge (1998)
15. Rich, E., LuperFoy, S.: An architecture for anaphora resolution. In: Proceedings of the second conference on applied natural language. Computational Linguistics, pp. 18–24. MIT Press, Cambridge (1988)