# Using 'Low-cost' Learning Features for Pronoun Resolution

Ramon Ré Moya Cuevas and Ivandré Paraboni

University of São Paulo, Escola de Artes, Ciências e Humanidades (USP / EACH)
Av. Arlindo Bettio, 1000, São Paulo, Brazil.
{fusion, ivandre} @usp.br

**Abstract.** We investigate a machine learning approach to Portuguese pronoun resolution. We presently focus on so-called 'low-cost' learning features readily obtainable from the output of a part-of-speech tagger, and we largely bypass deep syntactic and semantic analysis. Preliminary results show significant improvement in resolution precision and recall, and are comparable to existing rule-based approaches for the Portuguese language spoken in Brazil.

**Keywords:** Anaphora resolution, Machine Learning.

## 1. Introduction

The computational resolution of anaphoric expressions lies at the heart of a variety of NLP applications, including text understanding, Machine Translation, text summarization and many others. Although it has received a great deal of attention for many years now, anaphora resolution remains a computational problem yet to be overcome (Mitkov, 2002), and a challenge that is considerably increased if we speak of languages for which basic NLP resources (such as parsers, taggers or large corpora) are still under development, or may have only recently become available. This is the case, for instance, of Portuguese, one of the most widely-spoken languages in the world, and which still lacks somewhat behind as a relatively resource-poor language in NLP.

In this work we extend our previous investigation on learning approaches to Portuguese personal pronoun resolution in (Cuevas et. al., 2008.) In doing so, we focus on so-called 'low-cost' learning features, that is, we will limit the proposed solution to the knowledge readily obtainable from basic NLP tools such as part-of-speech taggers, and we will largely bypass deep syntactic or semantic analysis. In this sense, our work resembles the knowledge-poor approach in Kennedy & Boguraev (1996), which consists of a re-interpretation of the 'classic' algorithm proposed in Lappin & Leass (1994) using shallow rather than in-depth analysis. In addition to that, as we do not intend to explicitly write any anaphora resolution algorithms or rules (but rather induce them automatically) our work is mainly related to machine learning approaches such as Soon et. al. (2001), McCarthy and Lehnert (1995) and Ng & Cardie (2002). However, in discussing a possible 'low-cost' learning approach to Portuguese third person plural pronouns ("Eles/Elas"), we will focus more on the *choice* of learning features, and less on the *results* of a particular machine learning approach, which are to be discussed elsewhere.

The rest of this paper is structured as follows. Section 2 reviews previous work taken as the basis for our present investigation. Section 3 proposed an extended set of features for the problem at hand. Results of a standard decision-tree induction algorithm using the new features are presented in Section 4. Finally, Section 5 draws a number of comparisons with related work in Portuguese pronoun resolution and Section 6 describes our future work.

*22nd Pacific Asia Conference on Language, Information and Computation, pages 377–383*

## 2. Previous Work

As in Cuevas et. al. (2008), we will follow Soon et. al. (2001) and regard anaphora resolution as a machine learning *classification* task. Accordingly, a pronoun *j* and a potential antecedent term *i* may be classified as co-referent or not, that is, for each pair (*i, j*) in the text, we intend to label a binary class *coref* as being *co-referential* or *non co-referential*.

Positive instances of co-reference will consist of pairs (*i, j*) explicitly defined as co-referential in the training data by human annotators, and negative instances will consist of all pairs (*i, j*) in which *i* is an intermediate NP between *j* and its actual antecedent in the text.

For instance, the pronoun *j1* in the following text gives rise to one positive (*i1*, *j1*) and two negative ( (*i2*, *j1*) and (*i3*, *j1*) instances of anaphora. Analogously, pronoun *j3* also co-refers with *i1*[1], and pronouns *j2 and j4* both co-refer with *i2*:

> *Scientists*$_{\mathbf{i1}}$ *know that the phenomenon*$_{\mathbf{i2}}$ *occurs once every three to seven years*$_{\mathbf{i3}}$*: they*$_{\mathbf{j1}}$ *can detect when it*$_{\mathbf{j2}}$ *is coming, they*$_{\mathbf{j3}}$ *perceive when it*$_{\mathbf{j4}}$ *is going away.*

The starting point of the work in Cuevas et. al. (2008) was the Portuguese portion of an English-Portuguese-Spanish parallel corpus tagged using the PALAVRAS tool (Bick, 2000), comprising 646 articles (440,690 words in total) from the Environment, Science, Humanities, Politics and Technology supplements of the on-line edition of the "Revista Pesquisa FAPESP", a Brazilian journal on scientific news. Focusing on instances of third person plural pronouns (male) "Eles" and (female) "Elas", two independent annotators created a data set of 2595 instances of co-reference, being 483 positive and 2112 negative, with an average of 4.4 intermediate antecedents between each pronoun and the actual antecedent. About 10% of the positive instances were set aside with their negative counterparts for testing purposes. Thus, the test data comprised 234 instances and the reminder 2361 instances (being 435 positive or co-referential, and 1926 negative or non co-referential) became our training data. As we are still in the process of defining which precise features are applicable to the task at hand, our investigation is currently based on the training data set only, leaving the test data reserved for future use.

It was also shown in Cuevas et. al. (2008) that a simple set of syntactically-motivated features (based on distance, gender and number agreement) may achieve overall positive results in pronoun resolution (85.81% success rate) using C.4.5. ten-fold cross-validation decision-tree induction (Quinlan, 1993). However, this simple approach still suffers from low precision for the co-referential cases, making the resulting algorithm only partially useful for practical purposes. A more conservative (and possibly more reliable) analysis of these results focusing on the positive (i.e., co-referential) instances only shows a 70.5% score in F-measure. The following Table 1 summarizes those findings.

**Table 1:** Results from Cuevas et. al. (2008.).

| Class | Precision | Recall | F-measure |
|---|---|---|---|
| Co-referential | 0.572 | 0.910 | 0.703 |
| Non Co-ref. | 0.977 | 0.846 | 0.907 |

[1] In fact, pronoun *j3* co-refers with pronoun *j1* as well, although we presently do not deal with the resolution of full co-reference chains.

## 3. An Extended Set of 'Low-Cost' Learning Features

What kinds of learning feature may boost pronominal anaphora resolution? From the results in the previous section it is clear that additional features are needed to improve the system's ability to tell actual antecedents apart from all potential candidates. Thus, the first step in our present investigation was to extend the original set of features to gather as much information as possible about the anaphoric relations regardless of their usefulness to solve the problem at hand (which will be left to be 'learned' automatically.) However, as our ultimate goal is the induction of a Portuguese pronoun resolution algorithm based on existing - and easily accessible - Portuguese NLP resources, we shall limit our set of features to those based on the knowledge obtainable from the Portuguese tagger PALAVRAS (Bick, 2000.) More specifically, we have not defined any feature based on semantic knowledge other than what PALAVRAS may provide, or which may require full syntactic analysis.

Our extended set of features consists of 20 classes (plus the *coref* class to be learned), which are summarized in Table 2 below.

**Table 2:** An extended set of learning features for pronominal anaphora resolution given a candidate *i* and a pronoun *j*.

| Feature name | Description |
|---|---|
| *distance* | sentences between i and j. |
| *words_between* | number of words between i and j. |
| *same_sentence* | true if i and j occur in the same sentence. |
| *number_agreement* | true if i and j agree in number. |
| *gender_agreement* | true if i and j agree in gender. |
| *pronoun_type* | 1=personal ('eles/elas', or 'They'); 2=possessive ('deles/delas', or 'theirs'); 3=location ('neles/nelas', or 'in them' or 'on them'.) |
| *i_name* | true if i is a proper name. |
| *i_defined* | true if i is a definite description. |
| *i_demonstrative* | true if i follows a demonstrative. |
| *i_subject* | true if i is the sentence subject. |
| *i_direct* | true if i is a direct object. |
| *i_indirect* | true if i is an indirect object. |
| *j_subject* | true if j is the sentence subject. |
| *j_direct* | true if j is a direct object. |
| *j_indirect* | true if j is an indirect object. |
| *function_agreement* | true if j and i are both subject or object. |
| *is_hh* | true if i is a group of humans. |
| *is_org* | true if i is an organisation. |
| *is_inst* | true if i is an institution. |
| *is_civ* | true if i is a city, country, province etc. |

The feature *distance* counts the number of sentences between the pronoun and the candidate, under the assumption that an anaphoric relation becomes less likely as we move further away from the reference. Similarly, the *words_between* feature counts the number of words between pronoun and candidate, which may be particularly helpful for resolving *intrasentencial* anaphora, and so does (perhaps rather redundantly) the boolean feature *same_sentence*.

Given that Portuguese personal pronouns must always agree in number and gender with their antecedents, the features *number_agreement* and *gender_agreement* are expected to play a crucial role in the resolution of "Eles/Elas" references.

The feature *pronoun_type* accounts for the different pronoun usages most commonly found in our corpus: personal, possessive or locative. The features *i_name*, *i_defined* and *i_demonstrative* give additional information about the referring expression that represents the

candidate term: a proper name, a definite description or a description following a demonstrative pronoun (e.g., “that company”.)

In Cuevas et. al. (2008) the feature *function_agreement* was found to be unhelpful for anaphora resolution. However, this is not to say that subject/object information is irrelevant to our problem. On the contrary, such information is most likely essential to capture a number of syntactic constraints on pronoun resolution, especially considering that in-depth parsing information is not available. A possible reason why *function_agreement* was not useful in our previous work may be related to the excess of information that we attempted to convey as a single feature. Thus, in the presently extended set of features this information is split into six separate features (namely, *i_subject*, *i_direct*, *i_indirect*, *j_subject*, *j_direct* and *j_indirect*), from which we expect to derive the required syntactic constraints as originally intended, whilst allowing each individual feature to influence the solution independently.

The last four features (*is_hh*, *is_org*, *is_inst* and *is_civ*) are based on the semantic tags <*hh*>, <*org*>, <*inst*> and <*civ*> provided by PALAVRAS, and are intended to aid in the resolution of cases of anaphora in which there is no number agreement between antecedent and pronoun, as in, e.g., "The family" referred to as "They".

Finally, note that the above feature set is readily obtainable from a part-of-speech tagger such as PALAVRAS (Bick, 2000), in this sense corresponding to the ‘low cost’ aspect of our approach.

## 4. Testing

In order to select the most useful features for solving the problem at hand we started by taking the entire set of 20 learning features into account. Using C.4.5. ten-fold cross-validation decision-tree induction (cf. Quinlan, 1993) over the training data set alone, we confirmed the findings in Cuevas et. al. (2008) suggesting that the information conveyed by the *function_agreement* feature is not directly useful to our learning approach.

As for the additional features now under consideration, we manually tested several possible combinations to refine the resolution model. Speaking of the information about the anaphor (*j*), we found that *j_direct* and *j_indirect* did not improve resolution. This is largely explained by the fact that we focused on third person pronouns that do not occur in object position, that is, the syntactic function of the anaphor does not play a significant role in the resolution process[2].

Regarding the information about the candidate (*i*), four other superfluous features were identified: *i_indirect*, *is_org*, *i_name* and *i_demonstrative*. Once again, this was to be fully expected as, in a machine learning approach, we were not concerned with any linguistic investigation on how precisely pronoun resolution should be carried out, that is, one of the main goals of our investigation was precisely to determine which features are relevant or not.

The seven superfluous features (*function_agreement*, *j_direct*, *j_indirect*, *i_indirect*, *is_org*, *i_name* and *i_demonstrative*) were hence removed from the data and our test was re-run using the remaining 13 features. The following Table 3 summarizes our findings, using once again C.4.5. ten-fold cross-validation decision-tree induction (cf. Quinlan, 1993) over the training data set alone. The corresponding confusion matrix is shown in Table 4.

**Table 3:** Results from the extended set of features.

| Class | Precision | Recall | F-measure |
|---|---|---|---|
| Co-referential | 0.710 | 0.713 | 0.712 |
| Non Co-ref. | 0.936 | 0.935 | 0.936 |

[2] In principle these features are still relevant if, for example, we are to extend the existing approach to cover other kinds of reference phenomena.

**Table 4:** Confusion matrix.

| | True | False |
|---|---|---|
| True | 311 | 125 |
| False | 127 | 1830 |

The above confusion matrix is to be interpreted as follows: 2141 instances (being 311 co-referent and 1830 non co-referent) were correctly classified (89.47% success rate); 125 co-referent instances were misclassified as non co-referent (5.22%), and 127 non co-referent instances were misclassified as co-referent (5.31%.)

## 5. Discussion

At first glance, the present results are only marginally better than those achieved in Cuevas et. al. (2008). However, they do show improvement over our previous tests in the sense that they represent a better balance between precision and recall for positive instances of anaphora.

Regarding, existing work on Portuguese pronoun resolution, three of the best-known studies in the field are summarized as follows:

- Coelho & Carvalho (2005) describe an implementation of the Lappin & Leass algorithm (Lappin & Leass, 1994) for Portuguese third person pronoun resolution. The proposed algorithm was tested against 297 pronouns, achieving 35.15% success rate.
- Santos & Carvalho (2007) focus on an implementation of the Hobbs' algorithm (Hobbs, 1978) for Portuguese pronoun resolution. The test involved a set of 916 instances of non-reflexive pronouns in three linguistic genres, with accuracy rates ranging from 40.4% (texts on legislation) to 50.96% (magazine articles.)
- Chaves (2007) describes an implementation of the algorithm of R. Mitkov (Mitkov, 2002) for Portuguese third person pronouns. Results in this case range from 38% (novels domain) to 67.01% (newspapers articles) success rates.

A comparison between the best results achieved by these approaches and ours suggests that our present work is at least comparable to those. Moreover, being trainable from corpora, our work is in principle domain-independent, and much less prone to the wide fluctuations in results experienced by the above-mentioned studies.

On the other hand, it should be pointed out that when building the present training data, the annotators were selective in the choice of training instances of anaphora to be addressed. In particularly, our work does not include instances of reference to compound antecedents (e.g., "The boy and the girl" referred as "They"), which may partially explain the higher success rates[3].

Another important difference between learning and non-learning approaches to anaphora resolution is that in the former what counts as 'success' is simply the correct true/false labeling of the class 'coref', which is not the same as finding the right antecedent (as in the above mentioned non-learning approaches.) For example, our approach may successfully find the intended antecedent but, at the same time, mark a second candidate as co-referent as well, which may be correct (i.e., if they form a single co-reference chain) or not.

Bearing in mind these differences, the following summary in Table 8 is presented for illustration purposes only. Regarding our own work, we take a conservative view and show the F-measure score for co-referential cases only (cf. Table 6) and not the overall success rate of 89.47% since the data are heavily imbalanced (with on average 4.4 false antecedents for each pronoun.)

[3] To minor this difficulty, a separate annotation task is underway, in which a wider variety of reference phenomena will be taken into account to create a complementary test data set.

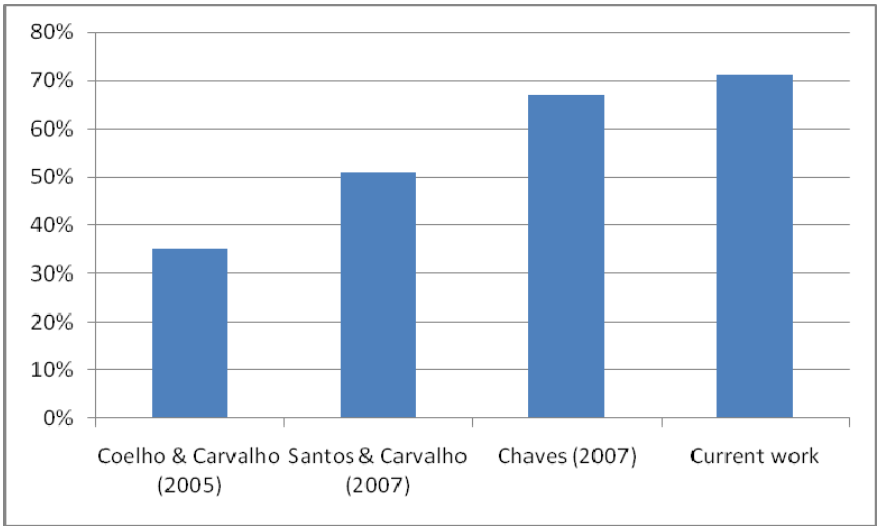


**Figure 1:** Maximum accuracy reported in previous anaphora resolution algorithms for the Portuguese language.

## 6. Conclusion

We have described an extension of previous work in Cuevas et. al. (2008) regarding a machine learning approach to Portuguese personal pronoun resolution. Using an enlarged set of features, our present results show improvement in resolution accuracy whilst avoiding the need for deep syntactic or semantic parsing information, which may not be easily obtainable for large-scale NLP projects involving the (Brazilian) Portuguese language.

We are now in the process of analyzing the remaining classification errors to define additional features to improve results even further. Among these, our approach may require information about adjunct and embedded expressions, as well as quantifiers and indefinite noun phrases usage. Since all the required information is (in principle) readily available from our tagged corpus, we expect to benefit from these additional features whilst keeping our knowledge acquisition costs low.

Once our set of features is stabilized and suitably tested, we intend to run our resulting pronoun resolution algorithm using the Portuguese portion of our parallel corpus as input, and use its output information to resolve their Spanish and English counterparts without any explicit knowledge about these languages. In doing so, we expect to improve the performance of an ongoing Machine Translation project for these three languages.

Finally, although in this work we have built our training data from a collection of third person plural pronouns only, we notice that our resulting algorithm should be capable of dealing with singular cases as well (i.e., “ele/ela” or he/she), and that should remain the case despite the fact that some of our current features (i.e., those conveying semantic ‘group’ information) are unlikely to play a role in the resolution of these cases. To make this point clear, a separate evaluation work on a different domain (namely, using a corpus of Brazilian newspapers articles) is underway, and will be described elsewhere once finalized.

## Acknowledgements

The authors acknowledge support by FAPESP (2006/03941-7, 2007/07356-4) and CNPq (484015/2007-9.)

## References


Bick, E. 2000. The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework. PhD Thesis, Arhus University.

Chaves, Amanda. 2007. A resolução de anáforas pronominais da lingual portuguesa com base no algoritmo de Mitkov. Msc. dissertation, University of São Carlos, São Carlos, Brazil.

Coelho, T.T. and Ariadne M.B.R. Carvalho. 2005. Uma adaptação de Lappin e Leass para resolução de anáforas em português. Anais do XXV Congresso da Sociedade Brasileira de Computação (III Workshop em Tecnologia a Informação e da Linguagem Humana – TIL 2005), São Leopoldo, Brazil, pp. 2069-2078.

Cuevas, Ramon Ré Moya, Willian Yukio Honda, Diego Jesus de Lucena, Ivandré Paraboni and Patrícia Rufino Oliveira. 2008. *Portuguese Pronoun Resolution: Resources and Evaluation*. 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008) Haifa, Israel. Springer LNCS vol. 4919, pp. 344-350. Springer-Verlag Berlin Heidelberg.

Hobbs, J. 1978. Resolving pronoun references. Lingua, vol. 44, pp. 311-338.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. 16th International Conference on Computational Linguistics (COLING-1996) Copenhagen, pp. 113-118

Lappin, S. and H. J. Leass. 1994. An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4), pp. 535-561.

McCarthy, J. F. and W. G. Lehnert. 1995. Using Decision Trees for Coreference Resolution. 14th International Conference on Artificial Intelligence IJCAI'1995.

Mitkov, Ruslan. 1999. Multilingual Anaphora Resolution. Machine Translation volume 14, numbers 3-4. Springer, pp. 281-299.

Mitkov, Ruslan. 2002. Anaphora Resolution. Longman.

Ng, Vincent and Claire Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp.104-111.

Quinlan, J.R.. 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Santos, D.N.A. and Ariadne M.B.R. Carvalho. 2007. Hobbs' algorithm for pronoun resolution in Portuguese. 6th Mexican International Conference on Artificial Intelligence, MICAI-2007, Aguascalientes, pp. 966-974.

Soon, Wee Meng et. al. 2001. A Machine Learning Approach to Correference Resolution of Noun Phrases. Computational Linguistics 27(4).