Coreference Resolution for Portuguese: Person, Location and Organization

Evandro B. Fonseca¹, Renata Vieira¹, Aline Vanin¹

¹Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br,
aline.vanin@ymail.com

Abstract. Coreference resolution is a process that consists in identifying the several forms that a specific entity may be referred to in a discourse. The automatic textual coreference resolution is a very important task in Natural Language Processing (NLP), because several others need its output. The present work shows a system for coreference resolution for Portuguese and its evaluation. The system solves coreference for three entity categories: Person, Location and Organization. These three categories were chosen because of their relevance for most NLP tasks, given that they represent concrete entities of common interest. Furthermore, they are the most explored categories in the related work. We also chose to work only with open source resources, whereas most of related works for Portuguese uses commercial software, which limits their availability and usability. The methodology is based on supervised machine learning. Based on a set of features the classification of noun phrase pairs as coreferent or non-coreferent is learned. The pairs are grouped later, thus building coreference chains.

Key-words: Coreference Resolution; Natural Language Processing; Named Entities; Machine Learning.

1. Introduction

Coreference resolution is a process of identifying the several occasions that the same entity may be referred to on a given discourse. In the sentence: "Natália has passed on the exam. The student is very happy with the news", we can say that "The student" is coreferent to "Natalia". The coreference resolution is an important task and also a challenge for the field of computational linguistics. And, in the case of Portuguese, this challenge is even greater. That is, the amount of resources for Portuguese is very limited when compared with the amount of available resources for other languages such as English. Nowadays, coreference is a well-known research subject. Annually, competitions like CoNLL (Conference on Computational Natural
Language Learning) [4] are performed, aiming to motivate the development of systems that solve coreference, focusing mainly on English. The CoNLL provides a corpus called Ontonotes [12], containing coreference marks (gold mentions). This corpus serves to measure the accuracy of the systems developed by the candidates.

On the Portuguese language field, there is HAREM [8]. Like CoNLL, HAREM is a shared evaluation task with the objective of promoting the research on NLP. In 2008, a task related to the named entity recognition was proposed for the first time. The HAREM provides a corpus with coreference annotations for the same purpose of CoNLL: evaluating the developed systems. The great contrast between these two corpora is their sizes. The Ontonotes corpus has 1.3 million of annotations, divided in several layers, like: syntactic layer, preposition layer, named entities, coreference and word sense disambiguation (WSD); in contrast, the HAREM [8] corpus has a little more than 290 thousand annotations. This gives an idea of the difference between the quantities of annotation for the two languages.

This paper are organized from the following form: In present chapter, is given a little introduction about coreference resolution; in chapter 2, we presents the main related works for the English and Portuguese; in chapter 3, we describes the features used in the model; in chapter 4, is describe our model for coreference resolution; and, in subsequent chapters, we present respectively: validation and conclusion.

2. Related Works

Reviewing the literature about coreference resolution, we found some works that are purely rule-based and others which use a more dynamic approach, based on machine learning. In CoNLL 2011 [4] (Conference on Computational Natural Language Learning) [10] presented their system purely based on rules for coreference resolution on English. Contradicting the meaning of the word "learning", [10] showed the efficiency of their system, ranking first in the 2011 CoNLL competition. The system, "Stanford's Multi-Pass Sieve Coreference Resolution System", purely deterministic, reached an efficiency of 57.79%. This efficiency was measured by the average of three performance metrics (MUC, B-CUBED and CEAFε), described in [12].

In 2012, in CoNLL, [6] presented the following strategy: a system of machine learning based on a perceptron algorithm. His proposal was based on two main techniques: latent coreference trees and entropy guided feature induction. The system has a few basic steps, such as: (a) Mention detection: for each text document a list was generated using the strategy of [13], containing the candidate mentions. The basic idea was to use all the noun phrases and additionally named entities. Verbs were not included as mentions. (b) Mention Clustering: in the mention clustering subtask, an instance of training (x, y) consists of an x group of mentions to a
document and y coreference groups. The structure of the perceptron algorithm learning for a given training set \( D = \{ (x, y) \} \) correct pairs of input/output.

(c) Coreference trees: to reduce the complexity prediction mentions problems, [6], the winners of the CoNLL competition in 2012, used trees to represent the grouping of terms that are coreferent, solving coreference in multiple languages. A coreference tree is a representation where the nodes are directed for mentions, and the arcs represent some relationship between coreferent mentions. The accuracy of the system, according to the metrics used by [PRA11], was 58.49% for Chinese, 54.22% to 63.37% for Arabic and English, getting an overall score of 58.69%.

For Portuguese, Silva [14] has proposed, in his master dissertation, a system of coreference resolution based on non-supervised machine learning. His system was divided into two stages: mention detection (noun phrases) and features and coreference chains detection. The first stage, mention and feature detection, has a text set as input. Newspaper texts that have the same subject were used. These texts were previously grouped, because the system does not have a clustering stage for identifying and extracting attributes of noun phrases. [14] used the syntactic analyzer PALAVRAS [1], the named entities recognizer Rembrandt [2], and the thesaurus TeP2.0 [11]. The second stage uses the output of the previous stage as input. With this information, the mention and coreference chains are grouped. The stage begins using a non-supervised machine learning method for the first grouping. After that, heuristic rules are applied in order of improving the generated chains. The results of this evaluation proved to be promising: 58.11% using MUC measure, and 60.07% using B-CUBED. Although a comparison with the [10]’s and [6]’s systems is not possible, due to difference in corpora, language and kind of entities processed, the proposed work by [14] had a significant contribution by addressing Portuguese domain.

Our system is based on supervised machine learning. Because of that, the model is divided in two steps. In the first step, the model must be trained. For this, we used the Summ-it corpus [3] to extract features and attributes. Through this, a training data set was created. In order to build the classifier, we used a decision tree, generated by SimpleCart algorithm [7]. The second step of our model is the coreference resolution phase: in this step, the system uses embedded resources for the feature extraction. Later, a feature vector is submitted to classifier, aiming to solve and group the coreference chains. Yet, our system has a few limitations, like solving coreferences only for proper names, kind of Person, Location and Organization. However, the great point on our system is that it solves coreferences for all kinds of plain texts, and not only for a specific corpus domain.

3. Classification Model

To build the classifier we selected features based on the literature review, mainly on conducted experiments by [16]. Soon et al. did an experiment in order to verify the impact of certain features of the correct coreference pair classification. The results show that certain features such String Match, Alias and Appositive, returns
good results. In addition to the more relevant characteristics, through conducted experiments, we feel the lack of a feature that set a Boolean value for a given distance between the noun phrases, according to Table 1.

<table>
<thead>
<tr>
<th>Features description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{P_StringMatch}</td>
<td>If an ‘x’ noun phrase contains ‘y’ noun phrase (Except stopwords)</td>
</tr>
<tr>
<td>Alias</td>
<td>If the words of the x noun phrase are abbreviations of any ‘y’ noun phrase word</td>
</tr>
<tr>
<td>\textit{M_Gênero}</td>
<td>If the noun phrases agree in Gender(Male/Female)</td>
</tr>
<tr>
<td>\textit{M_Número}</td>
<td>If the noun phrases agree in Number (singular/plural)</td>
</tr>
<tr>
<td>\textit{Categ_semântica_Igual}</td>
<td>If the entity categories (Person, Location or Organization) are equal. (False case the semantic category be unknown)</td>
</tr>
<tr>
<td>\textit{Categ_semântica_Diferente}</td>
<td>If the entity categories (Person, Location or Organization) are different. (False case the semantic category be unknown)</td>
</tr>
<tr>
<td>Distância&gt;5</td>
<td>If the number of sentences between one noun phrase and the other is greater than 5, returns true, false otherwise.</td>
</tr>
<tr>
<td>Distância&gt;10</td>
<td>If the number of sentences between one noun phrase and the other is greater than 10, returns true, false otherwise.</td>
</tr>
<tr>
<td>Distância&gt;15</td>
<td>If the number of sentences between one noun phrase and the other is greater than 15, returns true, false otherwise.</td>
</tr>
</tbody>
</table>

As a learning method we chose the SimpleCart algorithm. In our previous work [7] we tested other algorithms, such as \textit{Random Forest}, \textit{LBR} and \textit{Multilayer Perceptron}, which had a slightly higher performance than decision trees. However, since decision trees presented a performance that is compatible with other algorithms and present us an implementable structure, we adopted the classification model provided by Simple Cart and implemented it as an independent tool. In order to build this model and the experiments in [7], we use Weka\footnote{http://www.cs.waikato.ac.nz/~ml/} API
4. CORP

The CORP (Coreference Resolution for Portuguese) is a coreference resolution system for the Portuguese language, built using only open source resources. The objective of this toolkit is to help on several NLP tasks. According [9], the coreference resolution can provide significant gains for the NLP area. A good example is a relation extraction between named entities. Identifying the several forms that refer to the same entity on a certain text is make it possible to improve the relation extraction between named entities. For example, in the following sentence: “José da Silva resides near from Cidade Baixa, in Porto Alegre. Silva is in first year of his master degree on PUC-RS.”. Creating the coreference relation between the entities ‘José da Silva’ and ‘Silva’, makes it possible to induce a relation between ‘Silva’ and ‘Cidade Baixa’ (Silva resides on Cidade Baixa in Porto Alegre). We can also say that ‘José da Silva’ has a relation with PUC-RS (José da Silva is a student from PUC-RS). In the Figure 1, we can see the architecture of the system.

![Figure 1: CORP architecture.](image)

Given a *(plain text)*, the system executes the *(CoGrOO API)* [15] and extracts all noun phrases from the text. After that, the noun phrase pairs are generated. The pair generation is performed as follows: Each noun phrase makes pair with the next, for example: for the given noun phrases NP1, NP2, NP3 and NP4, the generated pair will be: {NP1-NP2, NP1-NP3, NP1-NP4, NP2-NP3, NP2-NP4, NP3-NP4} each noun phrase forms pair from the next, but never with the previous, like the method used by [5]. After the extraction of noun phrases and pair
generation, the system does the extraction of the features of these pairs, returning “true” or “false” for each attribute. The semantic category is given by auxiliary lists and Repentino^2. These lists contain proper names that help in identifying certain named entity categories. In Table 2 we can see the feature vector model.

Table 2: Feature Vector

<table>
<thead>
<tr>
<th>SN1</th>
<th>SN2</th>
<th>P.StrMatch</th>
<th>Alias</th>
<th>Gênero</th>
<th>Número</th>
<th>C.Sem. Igual</th>
<th>C.Sem. Dif</th>
<th>Dist&gt;5</th>
<th>Dist&gt;10</th>
<th>Dist&gt;15</th>
<th>Corref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlos</td>
<td>Nobre</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>?</td>
</tr>
</tbody>
</table>

After the feature extraction, the system submits the feature vector to the classifier, that based on these features, sets true or false on the “Corref” attribute (True for coreferent pair or false for not coreferent).

5. Evaluation

The CORP system evaluation was based on the MUC metric [17], using HAREM [8], a corpus available for Portuguese that has around 290.000 coreference annotations and named entity categories. Although it is not possible to make a direct comparison with the other systems, because of CORP restrictions, such as dealing only with proper names and specific kind of entities (Person, Location and Organization) , CORP obtained results at comparable level with related work. In Table 3, we can see the results of the related work using the MUC metric. The lower recall for CORP is due to the fact that this tool considers only proper nouns.

Table 3: Non-comparative results, using the MUC [17] measure.

<table>
<thead>
<tr>
<th>System</th>
<th>Language</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silva[14]</td>
<td>Portuguese</td>
<td>49.12%</td>
<td>45.9%</td>
<td>47.45%</td>
</tr>
<tr>
<td>CORP</td>
<td>Portuguese</td>
<td>77.97%</td>
<td>59.38%</td>
<td>67.42%</td>
</tr>
<tr>
<td>Fernandes[6]</td>
<td>English</td>
<td>75.91%</td>
<td>65.83%</td>
<td>70.51%</td>
</tr>
<tr>
<td>Fernandes[6]</td>
<td>Chinese</td>
<td>70.58%</td>
<td>52.69%</td>
<td>60.34%</td>
</tr>
<tr>
<td>Fernandes[6]</td>
<td>Arabic</td>
<td>49.69%</td>
<td>43.63%</td>
<td>46.46%</td>
</tr>
<tr>
<td>Lee[10]</td>
<td>English</td>
<td>59.3%</td>
<td>62.8%</td>
<td>61.00%</td>
</tr>
</tbody>
</table>

^2 http://www.linguateca.pt/repentino/
6. Conclusion

To conduct this research we performed a study about coreference resolution. Then, we identified techniques to build the model. Through these studies, we proposed an automatic coreference resolution method. Basically, the system receives a plain text document and returns an XML file, containing the coreference annotation. The system was evaluated through MUC metric, a very well consolidated metric in the NLP area. This work treats only proper names and entities from Person, Location and Organization. Despite this limited scope, this approach can be used as a basis for a more extensive model, which deals with coreference for all named entity categories, including other noun phrases and pronouns.

References


