Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Computational Intelligence Society, Vol. 12, Iss. 2, pp. 86-97, 2014
© Brazilian Computational Intelligence Society

COREFERENCE RESOLUTION IN PORTUGUESE: DETECTING PERSON, LOCATION AND ORGANIZATION

Evandro B. Fonseca[1]
Renata Vieira[1]
Aline A. Vanin[2]

[1]Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681 – FACIN, Faculdade de Informática
CEP 90619-900 Porto Alegre RS, Brasil

[2]Universidade Federal de Ciências da Saúde de Porto Alegre
Rua Sarmento Leite, 245, Prédio principal, Sala 412
Departamento de Educação e Informação em Saúde
CEP 90050-170 Porto Alegre, RS, Brasil

e-mails: evandro.fonseca@acad.pucrs.br, renata.vieira@gmail.com, alinevanin@ufcspa.edu.br

Abstract- Coreference resolution is a task of great relevance for Natural Language Processing area, given that the performance of many other tasks depends on the correct output of this type of system, especially the extraction of relationships between named entities. The present work aims at resolving coreference in Portuguese, focusing on the following categories of named entities: Person, Location and Organization. The proposed method uses supervised learning. To this end, the selection and implementation of features that assist in the correct classification are fundamental, since the classification model is built from this data.

Keywords- Coreference Resolution; Natural Language Processing; Named Entities, Machine Learning.

1. Introduction

Coreference resolution is the process of identifying the several forms an entity may assume in a text. In other words, this process consists of identifying certain terms and expressions that are related to a specific entity. As an example of coreference relation, we have the following: (1) “Schumacher$_{[i]}$ suffered an accident. The ex-pilot$_{[i]}$ is still in coma”. In this case, the noun phrase “The ex-pilot” is a coreference of “Schumacher”. In this work, we intend to study the automatic identification of named entities – such as “Schumacher” – and their coreference chains – which could be, in this case, “the Ferrari’s pilot”, “the ex-pilot”, “the skilled skier”, “the German pilot”, “the speed lover”, among others.

This study deals with coreference resolution focusing on specific named entities: Person, Location and Organization. Our main goal is to build and test a classifier for coreference resolution that is able to provide automatic coreference annotation from plain texts in Portuguese. We intend to perform this task using exclusively open source tools, such as a free corpus for testing and evaluating, a repository and a system for named entity recognition, as we will refer in this paper. Coreference resolution is a relevant task, though it is a great challenge for computational linguistics. While it is relatively easy to grasp coreference relations such as (2) “Jeff Mills” and (3) “Mills”, in which both NPs carry part of the noun “Mills”, it is a very complex task to relate the following noun phrases: (4) “A aranha” [The spider] and (5) “O aracnídeo” [The arachnid]. This happens because, in Portuguese, the gender of the nouns is different in each NP: in (4), the head of the NP is feminine and, in (5), it is masculine. When dealing with Portuguese, this challenge is even harder, because the quantity of resources for this language is limited when compared to all the resources available for other languages, such as English. The lack of resources for Portuguese may be seen, for example, in this comparison: Ontonotes (Pradhan et al., 2011) is a corpus for English language with around 1.3 million of words, distributed in five layers of annotation: Syntactic layer, Propositional layer, Named entities layer, Word Sense layer and Coreference layer. For coreference resolution there is a total of 131,886 mentions, 97,556 links and 34290 chains. While for Portuguese, Harem corpus (Freitas, 2010) has around 225 thousand words, distributed in three layers: coreference layer, relation between named entities layer (4803 marks) and semantic category layer (7847 recognized named entities). Abreu et al. (2013) propose the relation extraction among named entities in texts in Portuguese: the extraction of this type of relation has a considerable impact for the Natural Language Processing (NLP) field, given that this type of technique may improve the performance of many tasks. In this sense, the recognition of named entities aims at identifying, disambiguating and attributing a semantic category to these entities, such as Person, Location and Organization. This study intends to generate coreference chains from pure texts, that is, texts that are free of any annotation. These chains will be generated for specific named entities categories, contributing for the extraction of relations among entities through inferences. Gabbard et al. (2011) show that coreference resolution may provide meaningful gains for the relation extraction among named entities, since the coreference links may be useful for extracting sets of implicit relations.

Consider the following sentence: (6) “José da Silva mora perto do Centro, em Porto Alegre. O aluno está no primeiro ano de seu mestrado na PUCRS”. [José da Silva lives close to the Downtown area. The student is taking his first year of Masters at PUCRS]. When identifying and creating a coreference relation between “José da Silva” and “student”, it is possible to infer a direct relation between the entities “José da Silva” and “PUCRS” (in which José da Silva is a student at PUCRS). In other words, when we say that José da Silva is a student, it is possible to classify him as a person, as well as to say that he has relation with PUCRS.

The rest of this paper is organized as follows: Section 2 introduces the notion of coreference, the linguistic phenomenon in focus; Section 3 presents the related works, as well as the current state of the art and the detailed proposal of this work; in Section 4, we describe the resources that are used for this study; in Section 5, the features that were selected for the system are described; Section 6 gives details about the model of the classifier for coreferent pairs; finally, in Section 7, the conclusions and the future works are presented.

2. Coreference

Coreference is the relation between two words or phrases in which both refer to the same entity – in other words, they have the same referent, as in (7) and (8) below:

> (7) **João**$_{[i]}$ foi para a escola. **Ele**$_{[i]}$ está feliz com o início do novo ano.
> [João went to school. He is very happy with the beginning of the new year.]
>
> (8) **Felipe Massa** possui um excelente carro nesta temporada. **O piloto da Williams** alcançou a 9th no grid.
> [Felipe Massa has a great car this season. The Williams’ pilot reached the 9th position in the grid.]

In this type of relation, a grammatical substitute is used to refer to the denotation of a preceding word or expression. Perini (2009) claims that a noun phrase may be interpreted as a referential noun phrase (as well as an attributive or a qualifying noun phrase, though these aspects do not interest for our work). In this sense, interpreting an NP as referential denotes that an entity may be understood as existing or identifiable. For example, in (9) “**Aquele aluno** ganhou o prêmio” [“That student has got the prize.”], it is possible to identify the student who got the prize, while in (10) “**Apenas alunos** estão aptos a participar da competição” [“Only students are able to participate in the contest.”] (which is an attributive phrase) there is a description of a virtual – or general – entity, which may not exist.

According to Roncaratti (2010), these referring expressions are the ones that are used to refer to a linguistic or extra-linguistic entity, and they include: a definite expression, as in (11) “**O aluno** não fez a prova” [“The student did not take the test”]; an indefinite expression, as in (12) “**Um aluno** ainda não chegou” [“A student has not arrived yet”]; a personal pronoun, as in (13) “O estudante ainda não chegou, mas **ele** disse que virá” [“The student has not arrived yet, but he will come”]; a demonstrative pronoun, as in (14) “**Esse** aluno acaba de chegar” [“That student has just arrived”]; and a proper noun, as in (15) “**João da Silva** um grande aluno” [“João da Silva is a great student”], all of them mentioning a specific entity. Therefore, referring expressions are correlated to an entity; they identify the entities of the discourse, such as people, things, animals, places, events, processes, properties, predications etc. In general, noun phrases are the syntactic structures used for accessing a mentioned entity in a text.

The relation between a grammatical substitute and its antecedent is understood as a coreference relation (Juraksfy and Martin, 2000). Certain referent is represented in other points of the discourse as a “given” element, and it is not only known: the term that refers to that specific element implies both referenciation and coreferenciation (NEVES, 2011). When there is an identity between the antecedent and its anaphoric expression, there is an absolute coreference, though there are many cases of anaphor in which the total identification with its referent does not occur, such as in the example below:

> (16) “**Jornais**$_{[i]}$ do mundo todo noticiaram que **Gabriel García Márquez**$_{[j]}$ está hospitalizado desde a última semana. Em um comunicado, **o escritor Colombiano**$_{[j]}$ disse que **a imprensa**$_{[i]}$ deve preocupar-se com questões mais sérias.”
>
> [Newspapers around the world reported that Gabriel García Márquez is hospitalized since last week. In a statement, the Colombian writer said that the press should worry about more serious issues].

In (16), 'Jornais'[Newspapers] is the antecedent of 'a imprensa'[the press], and 'Gabriel García Márquez' is the antecedent of 'o escritor Colombiano'[the Colombian writer]. Interpreting such relation requires certain degree of world knowledge, which, in most cases, does not require great efforts for human readers. However, this may be quite challenging for machines. In this case, the more accurate the information on a specific entity is, the better the results will be.

> (17) "**Carlos Silva**$_{[i]}$ é um notável escritor. **Silva**$_{[i]}$ escreveu seu primeiro livro aos 12 anos de idade"
> [Carlos Silva is a notable writer. Silva has write your first book on 12 years age]

In (17) we got identic (IDENT) coreference, given be parcial matching [Carlos Silva] and [Silva] these examples is more easy to detect, but in some cases may be ambiguous.

> (18) "**Adalberto Portugal**$_{[i]}$ disse: em **Portugal**$_{[j]}$ as coisas são diferentes".
> ["Adalberto Portugual says: in Portugal the things are different"]

Note that in (18) Adalberto Portugal and Portugal are different entities, respectively person and location. Therefore is not corefent mentions.

Taking into account these assumptions, we may say that coreference resolution is a process which consists of identifying the several forms a named entity may assume in a text. In the sentence: (19) "Natália foi aprovada no vestibular. A estudante está muito feliz com a notícia" ["Natália was approved at SAT. The student is very happy with the news."], we may note that 'A estudante' ['The student'] is a coreference of 'Natália'. In this sense, identifying referring expressions and their structure in texts is fundamental for keeping the cohesion of a text, therefore giving more tools for better discourse interpretation. Moreover, when dealing with NLP, some of the tasks involving coreference resolution include the correspondence of the referent terms itself, information extraction, text summarization and machine translation.

3. Related work

Coreference resolution is a well-known natural language processing problem. In the literature, we find works that are only based in rules, and others that are based on machine learning. At the Conference on Computational Natural Language Learning (CoNLL 2011), Lee et al. (2011)) presented a system that is purely based on rules for coreference resolution in English. Even though the meaning of the word "learning" was not taken into consideration, Lee et al. showed that their system was efficient, and it got the first place in the competition. This system, "Sanforfd's Multi-Pass Sieve Coreference Resolution System", which is purely deterministic, reached an efficiency of 57.79%, and this was measured by the average rate among three performance metrics (MUC, B-CUBED and CEAFe), described in Pradhan et al. (2011).

In 2012, at CoNLL, Fernandes et al. (2012) presented the following strategy: a machine learning system based on a perceptron algorithm. This proposal was based on two main modeling techniques: latent coreference trees and entropy guided feature induction. The system has some basic steps, such as:

(a) Mention detection: a list for each document was generated containing the candidate mentions. In order to do so, Santos et al.'s (2011) strategy was used: the basic idea was to use all the noun phrases and, additionally, named entities of the text. Verbs were not included as mentions.

(b) Mention Clustering: in the subtask grouping of terms, a training instance (x, y) consists of a group of mentions x to a document and their coreference groups y. The structure of the perceptron algorithm learns for a given training set D = {(x, y)} of correct input / output pairs.

(c) Coreference trees: in order to reduce the problem of complexity of prediction references, Fernandes et al. (2012) used trees to represent the grouping of terms that are coreferent among themselves. A coreference tree is a tree whose nodes are directed to the mentions, and the arcs represent some relationship between coreferent mentions.

The task of CoNLL in 2012 was the resolution of coreference in three languages: English, Chinese and Arabic. According to Fernandes et al., their system, based on machine learning, may be easily adapted to different languages. In their experiments, small changes were necessary to resolve coreference in three different languages. The need for adjustments in the system was due to: (a) lack of input features for some languages; (b) different groups of tags used in Part-of-Speech (POS) of the corpora; and (c) the absence of a static list of specific pronouns of each language. Fernandes et al. were the winners of the CoNLL competition in 2012, solving coreferences in multiple languages as the proposed task. The accuracy of the system, according to the metric used by Pradhan et al., was 58.49% for Chinese, 54.22% for Arabic, and 63.37% to English, and the overall score was 58.69%.

For Brazilian Portuguese (PT-BR), Silva (2011) proposed a coreference resolution system also based on specific domains (Person, Location and Organization) using an unsupervised learning algorithm. This system is basically divided into two phases: (a) identification of terms (noun phrases) and their characteristics; and (b) identification of coreference chains. The first phase, identification of mentions (as NPs) and their characteristics, has a set of texts as input. Journalistic texts that deal with the same subject were used. These texts were previously grouped, since the system itself does not have a grouping step to identify phrases and extract attributes. Silva used the parser PALAVRAS (Bick, 2000), the named entities recognizer Rembrandt (Cardoso, 2008), the thesaurus TeP2.0 (Maziero et al., 2008) and the corpus CST-News (Maziero et al., 2010). The second phase, identification of coreference chains, receives as input the output of the previous stage. With this information, the grouping of entries in chains is performed. This phase starts with the use of a method of unsupervised machine learning for a previous grouping. After grouping, heuristic rules were applied in order to improve the quality of the generated strings. A set of tools available to Portuguese was used. The authors argue that if the availability of tools was as high as it is today for English, the results could even improve the results. Even though, the results of the evaluation of their system are promising: F-measure of 58.11%, using the MUC measure, and 60.07% using B-CUBED. Although no direct comparison can be established between Lee et al.'s and Fernandes et al.'s systems, due to differences in corpora, in languages and in types of entities treated, the work proposed by Silva made a significant contribution as it addressed this area for Portuguese.

Coreixas (2010) proposes a coreference resolution for Portuguese (PT-BR), with focus on categories of named entities. According to the author, works related to English have had successful results when using specific categories of entities. Based on these assumptions, Coreixas hypothesized that the use of specific categories of named entities has a positive impact on the task of coreference resolution, since each category has distinct and well-defined characteristics. As the categorization defines the field, the use of semantic information as a support tool in the coreference resolution process becomes more feasible. Coreixas' system was based on machine learning, categorization of named entities such as Person, Organization, Location, Work, Thing and Other (from the corpus of HAREM (Freitas et al.,2010)), the parser PALAVRAS (Bick) and the resource from Summ-it corpus (Collovini et al., 2007). Coreixas compares two versions of the system, namely: Baseline and "Recorcaten" (REsolução de CORreferência por CATegorias de ENs, meaning coreference resolution by named entities categories). The first version aimed at generating pairs of phrases without considering the categories of NEs. The second generates pairs considering these types of entities. As a contribution, through experiments with both versions, Coreixas showed that the use of categories of entities provided an improvement in the percentage of correct answers to determine whether a pair is anaphoric or not. Also, it showed the importance of world knowledge for this line of research, given the fact that some categories, such as Event and Organization, did not show a satisfactory return on the classification of coreferent pairs. This happens because the process of disambiguation was not performed correctly, thus emphasizing the importance of databases with synonyms, such as Wordnet (Miller, 1995), to complement and support the resolution of coreference. This work has limitations: (a) the size of the corpus used in the experiments is not very big (it has only fifty texts); (b) there are not many resources for Portuguese; (c) according to Coreixas, the parser presents several problems of annotation.

As previously mentioned, there are many works within the context of coreference resolution, but these are mostly learning models tested for a specific corpus, aiming at calculating its precision or recall. Our work aims at building a classifier and a coreference resolution system for Portuguese (PT-BR), using only open source resources. Initially, outperform the state of the art results is not our main concern. Our goal in this stage is to minimize the false positives, due to the fact that in coreference resolution stage, the quantity of negative pairs will be much greater(approximately 30 negative pairs for each positive) In this paper, we present the first stage of this process. That is, the classifier and features implementation. Unfortunately, most of the works for Portuguese refer to systems that were evaluated for a specific annotated corpus but do not provide a off the shelf tool. Our work aims at implementing a coreference system that solves coreference for plain texts.

Table 1: State of the art, non-comparative results

System	type	Language	Corpus	Evaluation Metric	Precision	Recall	F-measure
Lee	Rule-based	English	Ontonotes	MUC B-CUBED CEAFe BLANC	59,3% 69,0% 46,8% 76,6%	62,8% 68,9% 43,3% 71,9%	61,0% 68,9% 45,0% 74,0%
Fernandes	Supervised Machine Learning	English	Ontonotes	MUC B-CUBED CEAFe	75,91% 77,69% 43,17%	65,83% 65,79% 55,0%	70,51% 71,24% 48,37%
Fernandes	Supervised Machine Learning	Chinese	Ontonotes	MUC B-CUBED CEAFe	70,58% 80,57% 37,88%	52,69% 62,99% 53,75%	60,34% 70,70% 44,44%
Fernandes	Supervised Machine Learning	Arabic	Ontonotes	MUC B-CUBED CEAFe	49,69% 72,19% 46,09%	43,63% 62,70% 52,49%	46,46% 67,11% 49,08%
Silva	Non-supervised Machine Learning	Portuguese (PT-BR)	CST-News	MUC B-CUBED	49,12% 71,29%	45,9% 63,31%	47,45% 67,07%
Coreixas	Supervised Machine Learning	Portuguese (PT-BR)	Summ-it	J48 Classifier (Positive class)	78,1%	54,8%	64,4%
Our System	Supervised Machine Learning	Portuguese (PT-BR)	Summ-it	Simple Cart Classifier (Positive class)	76,8%	88,9%	82,4%

In Table 1, we see that most of the work uses metrics for comparing correference chains, as MUC, CEAFe, BLANC and BCUBED. Note that a direct comparison between these systems is not possible due to the fact that each system has specific domains, languages and scopes. Coreixas and our system were evaluated as classifiers of correfent pairs.

4. Resources

In this section, we list the resources used in the construction of our model of coreference resolution: Corpus Summ-it, used for training and evaluation; Repentino, NERP-CRF and auxiliary lists, for named entity recognition in Portuguese; and Weka, a toolkit used to deal with algorithms in machine learning.

Corpus Summ-it

Summ-it (Collovini et al.) is a corpus consisting of fifty journalistic texts from the Science section from the newspaper Folha de São Paulo, taken from the corpus PLN-BR (Bruckschen et al., 2008). Each document corresponds to a text file (ASCII) with size between 1 kbyte and 4 kbytes (127-654 words). The texts were annotated with syntactic, coreference and rhetorical structure information. Summ-it also includes summaries constructed manually and automatically. The corpus has a total of 560 coreference chains with an average of 3 members (noun phrases for each chain). The more large chain has 16 members (noun phrases). The parser PALAVRAS (Bick) was used to process the corpus. In order to improve the visualization of the information extracted from the parser, the generated file was divided into three other files: a file with the information of tokens consisting of the token and its corresponding ID; another file with the information of phrases, that is, information defining the token ID of the beginning and of the end of the phrase; and another with the syntactic-semantic information associated with the token ID. The files are in XML format. The corpus Summ-it has had an important role in the training and the validation of the classification model.

Repentino

The Repository for Recognition of Named Entities, Repentino (Sarmento, 2006), is a public resource that contains an average of 490,000 samples of named entities, i.e., it is a large list containing several proper nouns, such as names of people, places, chemical substances, organizations, among others. The examples of entities, stored in Repentino, are divided into several categories, each of which containing several sub-categories in a tree structure, thus ensuring a fair organization of such examples.

NERP-CRF

NERP-CRF (Amaral, 2013) is a System for Recognition of Named Entities using Conditional Random Fields for Portuguese. That is, NERP-CRF is a classifier, as it is a named entities identifier. Given a text or a set of texts, the system may process them. The output is all named entities recognized by NERP-CRF, including their categories, such as: Person, Location, Organization, Event, among others.

Auxiliary lists

Both the resources used for the purpose of tagging entities (NERP-CRF and Repentino) have limitations within the context of search used. NERP-CRF system has an accuracy rate of 83.99%. This means that not all entities are classified correctly. In the case of Repentino, the problem is ambiguity. For example, when searching the entity 'Amazon', Repentino may identify both as in the categories of 'Location' or as 'Organization' – respectively, 'Amazon' and 'Bank of Amazon'.

Thinking about these two limitations, the idea was to get the entities that were in the NPs in both resources and align their results. Thus, the output becomes more reliable and accurate. When both resources return the same result, it is assumed that the category of named entity was measured correctly, but the results of these systems do not always agree. Then, three lists were created, one for each category of entity: Person, Location and Organization. When the resources return different results, these three lists are run in order to check the occurrence of these terms, aiming at labeling correctly the category of the entity in focus. These lists contain common and proper nouns, which help to identify the type of named entity. The terms were extracted from Wikipedia[1]. As an example, the list "Person" has commonly used professions and people's names, like 'agronomist', 'lawyer', 'engineer ', 'Diego' , 'João' , 'Aline', 'James' etc. The "Local" list has some names as cities, for example, and some common nouns, such as 'square', 'beach', 'city', 'hill', 'street', 'district', 'avenue', 'river', 'lagoon' etc. The "Organization" list uses proper nouns of known companies and common nouns, such as 'institute', 'agency', 'organization', 'ONG', 'party', 'trade' etc. These auxiliary lists are only used when the correlation does not exist between Repentino labels and NERP-CRF labels.

Weka

Weka (Boukckaert et al., 2013) is a collection of machine learning algorithms for data mining tasks. It has resources for pre-processing, classifying, clustering, visualizing, among others. Its implementation is made in Java language, which is portable. Therefore, Weka can be run in several platforms, taking advantage of the benefits of an object-oriented language. Weka also has an API that can be run directly from the command line.

5. Selection of features

The selection of features was based on the study of the literature on the topic, and we based this task mainly on the results obtained in Soon et al.'s (2001) experiments. These authors conducted an experiment in order to verify the impact of certain features in the correct classification of coreferent pairs. As a result, the authors found that some features, such as String Match and Alias, have a significant return on the correct classification of pairs. In addition to the most relevant features that exist in other works, experiments showed the need of adding new features, as we can see in Table 2.

[1] Available at: http://pt.wikipedia.org/wiki/. Last access: 15/03/2013.

Table 2: Description of the *features*

P_StringMatch	If at least one String from NPx is contained in NPy (except stopwords).
Alias	If one of the words from NP1 is acronym of NP2.
Gender	If the phrases agree in gender (male/female).
Number	If the phrases agree in number (singular/plural).
Semantic_Categ_Eq	If the categories of entities (Person, Location or Organization) are equal.
Semantic_Categ_Dif	If the categories of entities (Person, Location or Organization) are different (that is, when the system cannot assess a category, both Categ_igual and Categ_Dif are *false*).
Distance>5	If the quantity of sentences between a phrase and another is bigger than 5, the output is *true*, and it is *false* otherwise.
Distance>10	If the quantity of sentences between a phrase and another is bigger than 10, the output is *true*, and it is *false* otherwise.
Distance>15	If the quantity of sentences between a phrase and another is bigger than 15, the output is *true*, and it is *false* otherwise.

The construction of pairs of NPs was based on coreference information contained in the Summ-it corpus. Some attention was paid in the implementation of the features, in order to calibrate the precision and recall of each one. In the case of the feature Parcial_String_Match, all the stopwords were removed from the NPs. Then, the similarity is calculated using the Jaccard coefficient (Manning and SchützeH, 1999) to verify how similar a string is to another. The calculation of similarity returns a number close to '1' for identical strings, or numbers close to '0' to completely different strings. After testing, it was noted that there were some words that were quite similar, but they often returned the value ‘false’ because they had little variation, such as "profissional" and "profissionais" [both translated as “professional”, respectively in the singular and in the plural forms]. Through preliminary experiments, we defined the threshold '0.8', given the positive matching when the return is greater than or equal to '0.8'. Thus, we had better results by minimizing false positives and increasing range and accuracy. For the features Gender and Number, information was extracted from the Summ-it corpus (but this may also be acquired by means of LX-Tagger (Branco and Silva, 2004), a free Part-of-Speech tagger for Portuguese).

To the alias feature, as well as to Parcial_String_Match, similarity calculation was used. For each word in an NP (excluding stopwords), if the first letter is capitalized, it is selected and stored. As a result, there is a string with the initials of the words with and without period '.', as in the example: NP1 = "National Institute for Space Research, INPE" and NP2 = "INPE". The acronyms generated by the feature will be "INPEI" and "INPEI". After this step, the acronyms generated are compared with each word of the NP2 by calculating similarity. Note that “INPEI” is not exactly equal to the NP2 “INPE”, but the similarity calculation will give a very close result of "1", indicating that the strings are quite similar. Thus, it is concluded that SN2 is acronym of SN1.

For the feature Semantic_Categ, three resources were used: Repentino, NERP-CRF and auxiliary lists, which aim at identifying and assessing Location, Person and Organization labels for the pairs of phrases. When the system cannot define a category to certain entity, the value ‘false’ is attributed to these features: ‘Semantic_Categ_Eq’ and ‘Semantic_Categ_Dif’. This way, the learning algorithm may differentiate the entities with categories that are not distinguished from the other ones.

The implementation of the features Distance_5, Distance_10 and Distance_15 was based on the premise that the farther a sentence is from the other in a text, the smaller the chances are of these being coreferent. This feature uses the original text as a parameter, since it is free of annotations, and two NPs. Through the two NPs, the *feature* forms a regular expression, capturing from the text the entire existing excerpt between NP1 and NP2. Thus, the number of sentences is counted, considering the punctuation of the text (as ‘.’, ‘!’ and ‘?’). As a result, a value ‘true’ or ‘false' is set for each of these three features. Or, in cases in which there is more than one identical phrase in the text, the feature would never have the correct output, and, in this case, it returns ‘?’.

6. Generating a classifier for coreferent pairs

One of the main challenges of this work is the generation of candidate pairs to coreference. As the idea is to generate a model for solving coreference based on machine learning, the first step of this work is to train a classification model, which receives pairs of phrases, and, through its characteristics, it can predict whether these pairs are coreferent or not. The first task is relatively simple, because the coreference information contained in the training corpus should be organized and submitted to machine learning.

Once the model is generated, the system should be able to predict which pairs are coreferent and which are not through numerous pairs of positive and negative phrases – at this point, we are interested in knowing which pairs are, in fact, coreferent. As the amount of negative pairs is much larger, it is natural (and it is also a problem) to have negative pairs classified as positive and positive pairs considered negative ones. This occurs because there are no features (except Parcial_String_Match) that are crucial to the point of classifying certain NP pairs with a high accuracy rate. To solve this problem of misclassification, several heuristics have been created to reduce the amount of negative pairs; however, none were significant to the point of drastically reducing the number of negative pairs without reducing the number of positive pairs. An alternative method was to limit the scope of the system, taking into account only pairs that have proper nouns in both phrases.

Several classification models were generated before choosing one. After pre-processing, we obtained 3855 negative pairs and 394 positive pairs, still a high number of negative examples. We used the oversampling technique (Chawla et al., 2002) in the positive pairs until they reached an acceptable amount to generate the model, thus preserving all the negative pairs. The amount of negative pairs was maintained, not to take the risk of losing important characteristics that may exist in some pairs. As a result, 5431 pairs were used, 1576 positive (4x394) and 3855 negative. Then, on the basis of the features described before, we achieved the results shown in Table 3. In Table 3, we can see the results of the experiments carried out with some classification algorithms, aiming at measuring the performance according to different types of classifier algorithms. Obviously, all existing experiments in Table 3 were performed with the same data set. We also used the same set of Boolean features, since we could not discard the decision tree options, as we needed a clear model for our implementation. In Table 3, the classes P and N, respectively, represent positive and negative classes.

Table 3: Accuracy of the classifiers

-	Type	Precision P	Recall P	F-Measure P	Precision N	Recall N	F-Measure N	Accuracy
SimpleCart	Decision tree	76,8%	88,9%	82,4%	95,1%	89,0%	92,0%	88.97%
BFTree	Decision tree	76,4%	88,9%	82,2%	95,1%	88,8%	91,9%	88,82%
REPTree	Decision tree	75,4%	88,9%	81,6%	95,1%	88,1%	91,5%	88,36%
J48	Decision tree	76,3%	88,9%	82,1%	95,1%	88,7%	91,8%	88,78%
Random Forest	Decision tree	76,7%	89,8%	82,8%	95,5%	88,9%	92,1%	89,15%
LBR	Lazy	79,5%	87,3%	83,2%	94,6%	90,8%	92,7%	89,79%
NaiveBayes	Bayes	76,8%	84,1%	80,3%	93,3%	89,6%	91,4%	88,01%
Multilayer Perceptron	Neural network	79,1%	89,8%	84,1%	95,6%	90,3%	92,9%	90,16%

Table 3 shows that some algorithms, such as *RandomForest*, *LBR* and *Multilayer Perceptron*, had a slightly higher performance. However, we can see that the SimpleCart, which generates an implementable model, presented a performance that is compatible with these algorithms. We chose the SimpleCart algorithm because it presented the best results among decision trees. Also the model it generates can be implemented directly on the source code, discarding secondary resources like a Weka API (one our main concerns was to generate an embedded open source resource). We cannot really compare our results with Coreixas' model, because we deal only with proper nouns in both noun phrases, whereas her model includes all types of noun phrases along with proper nouns. Our system deals with a simplified version of the correference problem. Having said that, we consider that our results are promising: 88,97% of global accuracy against Coreixas' 70,26%. In the future, we intend to remove this restriction, solving coreference for all noun phrases.

In Table 4, we can see the confusion matrix of the generated model: 1401 positive pairs were correctly classified (true positives), 175 positive pairs were classified as negative (false negatives), 3431 negative pairs were classified as negative (true negatives) and 424 negative pairs were classified as positive (false positives).

Table 4: Confusion Matrix of SimpleCart.

	Class	
-	**P**	**N**
P	1401	175
N	424	3431

In Figure 1, we can see the decision tree generated by SimpleCart algorithm. We can see that the most relevant feature is a Parcial_String_Match (root of the tree). For each node, we implemented an "if" and an "else" directly on the source code. Our tree receives a feature vector (Table 5) and uses its nodes to classify it. The output is "true" or "false": "true" for coreferent pairs, false for non-coreferent pairs.

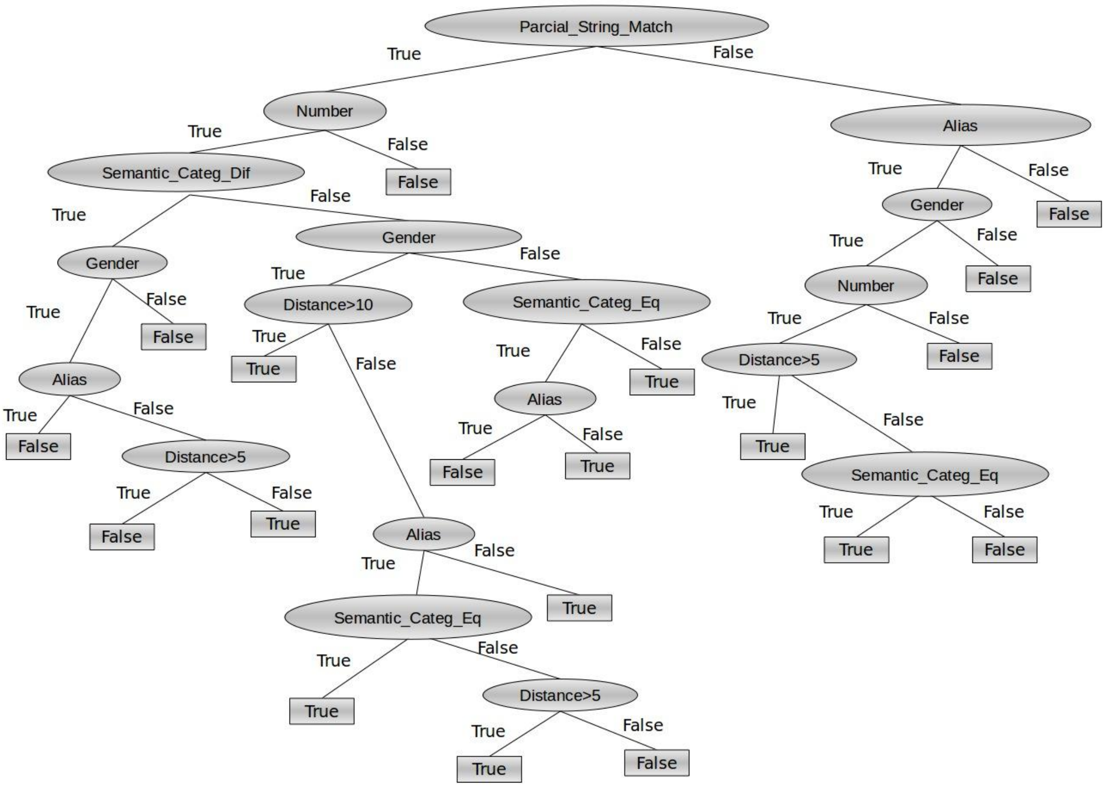

Figure 1: Decision tree generated by SimpleCart algorithm.

In table 5 we can see an example:

Table 5: Example for the sentence: "O humorista **Carlos Nobre** fará uma apresentação... **Nobre** informa que..."

Noun Phrases		**Feature Vector**								
NP1	**NP2**	**P_string Match**	**Alias**	**Gender**	**Number**	**Sem_Categ _Eq.**	**Sem_Categ_ Dif.**	**Dist> 5**	**Dist> 10**	**Dist> 15**
Carlos Nobre	Nobre	True	False	True	True	True	False	False	False	False

For the example presented in Table 5, we got the following path on the decision tree: Parcial_String_Match (True) >> Number (True) >> Semantic_Categ_Dif (False) >> Gender (True) >> Distance>10 (False) >> Alias (False) >> [TRUE].

7. Conclusion and Future work

In this paper, we described the generation of a model for building a system which is able to resolve coreferences focusing on categories of named entities. Although there are many related works, few studies focused on specific categories of entities. The use of specific categories of named entities have a positive impact on the task of coreference resolution, as each category has distinct and well defined features (Coreixas).

As any study, this proposal has limitations, such as restricted availability of free resources for Portuguese, as well as the low quality of these resources. As a contribution, this paper gives a classifying model for coreference resolution, as well as a system for coreference resolution for Portuguese. As future work, our aim is improving this system, expanding its scope to other semantic categories within Harem corpus, as suggested by Freitas et al.

Using the SimpleCart decision tree, we implemented a coreference resolution tool[2] with a limited scope – it deals only with proper nouns in the categories of Person, Location and Organization. On the positive side, it resolves coreference from pure texts. It is based on Cogroo (Silva, 2013) for NP and proper noun identification, Repentino and auxiliary lists (see Section 4) for NE classification, in the model described in this paper.

We had a high precision for Class P (positive), 76.8%, and for Class N (negative), 95.1%. The accuracy reached 88.9%. The precision of class P is what we intend to improve, without losing the recall.

Acknowledgements

The authors acknowledge the financial support of CNPq, CAPES and Fapergs.

References

Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013) **A review on Relation Extraction with an eye on Portuguese**, Pages 1-19 In: Journal of the Brazilian Computer Society. Available at: http://link.springer.com/article/10.1007%2Fs13173-013-0116-8#page-1

Amaral, D., (2013) **Reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**, Dissertação de mestrado, Pontifícia Universidade Católica do Rio Grande do Sul. Available at: http://repositorio.pucrs.br/dspace/handle/10923/5772

Bick, E., (2000) **The Parsing System "Palavras" - automatic grammatical analysis of portuguese in a constraint grammar framework**, Tese de Doutorado, Department of Linguistics, University of Århus, DK.

[2] The system is available at www.inf.pucrs.br/~linatural/corp.html

Boukckaert, R., Frank, E., Hall, M., Kirkby, K., Reutemann, P., Seewald, A. and Scuse, D., (2013) **Weka Manual for version 3.6.9**, The University of Waikato. Available at: http://www.cs.waikato.ac.nz/ml/weka/

Branco and J. Silva. (2004). **Evaluating Solutions For The Rapid Development Of State-Of-The-Art Pos Taggers For Portuguese**. In Proceedings of the 4th International Conferenceon Language Resources and Evaluation (LREC2004), pages 507–510. ELRA.

Bruckschen, M., Muniz, F., Souza, J., Fuchs, J., Infante, K., Muniz, M., Gonçalves. P., Vieira, R. and Aluísio, S., (2008) **Anotação Linguística em XML do Corpus PLN-BR**, USP. Available at: http://www.nilc.icmc.usp.br/nilc/download/Nilc_TR_08_09.pdf

Cardoso, N. (2008) **REMBRANDT – Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto**, chapter 11, pages 195-211. Linguateca, 1 edition. ISBN 9789892016566.

Collovini, S., Carbonel, T., Fuchs, J., Coelho, J., Rino, L. and Vieira, R. (2007) **Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática**. In: V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC, Rio de Janeiro. Available at: http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=11082

Chawla, N.V., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P. (2002), **Smote Synthetic Minority Oversampling Technique**. In Journal of Artificial Inteligence Research, 16:321-357

CoNLL2011, **Conference on computational natural language learning**, Available at: http://conll.cemantix.org/2011/. Access on: 05/08/2012.

Coreixas, T. (2010) **Resolução de Correferência e Categorias de Entidades Nomeadas**, Dissertação de Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul, FACIN. Available at: http://repositorio.pucrs.br/dspace/handle/10923/1567

Fernandes, E., Santos, C. and Milidiú, R. (2012) **Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution**, Conference on computational natural language learning. Available at: www.aclweb.org/anthology/W12-4502

Freitas, C., Mota, C., Santos, D., Oliveira, H. and Carvalho, P. (2010) **Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese**, Linguateca, FCCN. Available at: http://www.lrecconf.org/proceedings/lrec2010/pdf/412_Paper.pdf

Gabbard, R., Freedman, M. and Weischedel, R. M. (2011) **Coreference for Learning to Extract Relations: Yes, Virginia, Coreference Matters**. In: Proceedings 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 288–293, Portland, Oregon.

Jurafsky, D. and Martin, J**. Speech and language processing**. In: [S.l.]: Alan Apt, 2000. cap. Discourse, p. 670-718.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. (2011) **Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task**, Conference on computational natural language learning.

Manning C. D. and SchützeH. (1999) **Foundations of Statistical Natural Language Processing**, The MIT Press Cambridge, Massachussetts London, England, computational linguistics-Statistical methods. ISBN 0-262-13360-l, Page 299.

Maziero, E. G., Pardo, T. A. S., Felipo, A. D. and Silva B. C. (2008) **A Base de Dados Lexical e a Interface Web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil**. In WebMedia, Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, p 390-392, doi:10.1145/1809980.1810076

Maziero, E. G.., Jorge, M. L. C. and Pardo, T. A. S. (2010) **Identifying Multidocument Relations.** In: 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS, Funchal/Madeira. Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS. pages 60-69.

Miller, G. (1995) **WordNet: A Lexical Database for English.** In: Communications of the ACM Vol. 38, No. 11: 39-41.

Available at: http://dl.acm.org/citation.cfm?id=219748

Neves, M.H.M.(2011) **Texto e gramática**. São Paulo: Contexto.

Perini, M. A.(2009) **Gramática descritiva do Português**. 4. ed. São Paulo: Ática.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R. and Xue, N. (2011) **CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes**, CoNLL Shared Task.

Roncarati, C. (2010) **As cadeias do texto: construindo sentidos**. São Paulo: Parábola Editorial.

Sarmento L., Pinto A. S. and Cabral L.,(2006) **REPENTINO – A Wide-Scope Gazetteer for Entity Recognition in Portuguese**, In: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil. Vol 3960, Pages 31-40, doi: 10.1007/11751984_4

Santos, C. and Carvalho, D. (2011) **Rule and Tree Ensembles for Unrestricted Coreference Resolution**, 15th Conference on Computational Natural Language Learning.

Silva, J. (2011) **Resolução de Correferência em Múltiplos Documentos Utilizando Aprendizado Não Supervisionado**, Dissertação de Mestrado, USP.

Silva W. D. C. M. (2013) **Aprimorando o Corretor Gramatical CoGrOO**, Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo.

Soon, W., NG, H. and Lim, D. (2001) **A Machine Learning Approach to Coreference Resolution of Noun Phrases**. Computational Linguistics Vol. 27, No. 4, Pages 521-544, doi:10.1162/089120101753342653.