CORP: Coreference Resolution for Portuguese

Evandro Fonseca[1], Renata Vieira[1] and Aline Vanin[2]

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

[1]Pontifícia Universidade Católica do Rio Grande do Sul
[2]Universidade Federal de Ciências da Saúde de Porto Alegre

Abstract. This paper describes CORP, an open source, off-the shelf noun phrase coreference resolver for Portuguese with a web interface.

1 Introduction

We are building an open-source off-the-shelf system, which solves Portuguese noun phrase coreference, using plain texts as input. Our tool goes beyond basic syntactic heuristics, such as string matching, copular, juxtaposition. We consider semantics. In other words, string matching heuristics serve to deal with cases such as "Miguel Guerra" and "Guerra", in which both NPs share some identical part. Copular constructions are used to link two mentions as in "Miguel Guerra is the agronomist". Juxtaposition refers to cases of appositive constructions such as "Miguel Guerra, the agronomist". So far, it may seem a simple problem, but refined syntactic knowledge must be taken into account even in those cases. For instance, we do not want to find that "mushrooms found in Brazil" is coreferent with "mushrooms found in France". In other situations, establishing a coreference relation is even more difficult. In cases such as "the boy" and "the kid", there is a semantic relation which is usually part of the readers' common sense knowledge. We are currently dealing with this sort of problem. The current version of the system is available through a web interface and is detailed below.

2 CORP Architecture

In this research we are using what is currently available for pre processing tasks. As we are developing a system in Java, we have used Java based open source tools such as Cogroo [2] and OpenNLP[1] . OpenNLP provides POS tagging and named entities recognition, while Cogroo provides noun phrase chunks and shallow structure. For our studies, we use the correference annotated Summ-it corpus [1]. Our system is an adaptation of the model proposed in [5]. We adapted and implemented a set of modules. The first two correspond to noun phrase extraction and filtering. The other modules are used to link two mentions if the conditions established by linguistic rules are satisfied. These modules are described in detail in [3]. Recently, we added two sematic modules (Hyponymy and Synonymy) based on the relations provided by ONTO-PT [6]. An experiment using semantic knowledge is reported in [4].

[1] http://opennlp.apache.org/

3 CORP - Web Interface

A demonstration of the system is available at “http://ontolp.inf.pucrs.br/corref/”. The interface is intuitive and contains: a upload button, to submit the text; “Limpar texto” to clear input text and the output and three example buttons, containing samples of previously processed texts (Figure1). When submitting an input text, the system returns its coreference chains.

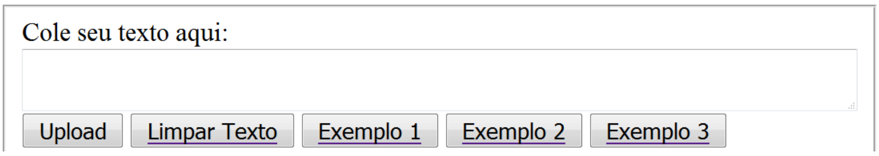
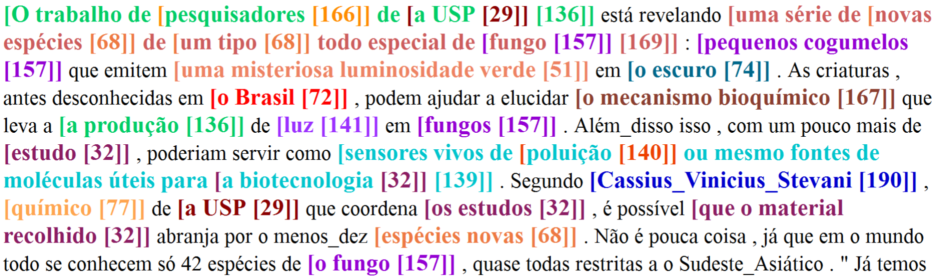


Fig. 1. CORP Web Interface



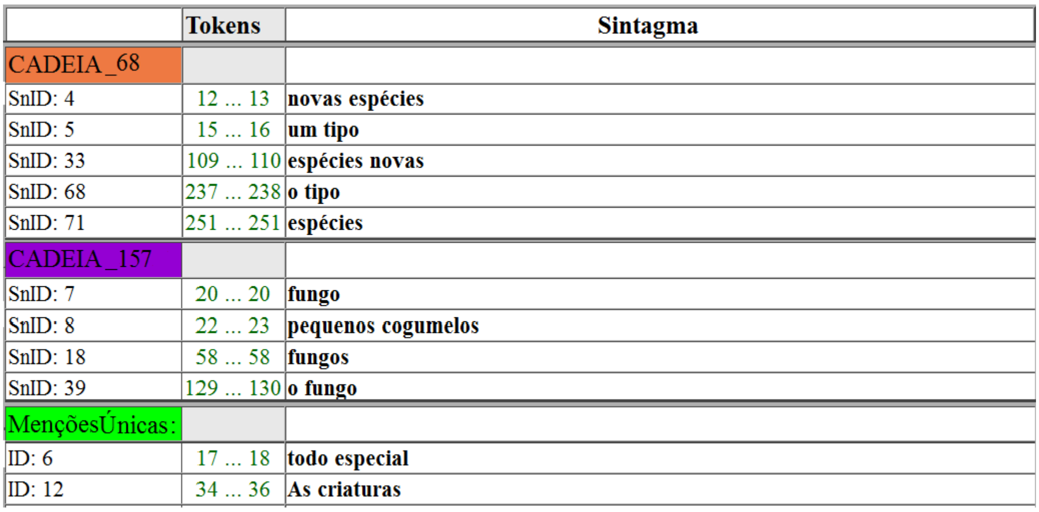
[O trabalho de [pesquisadores [166]] de [a USP [29]] [136]] está revelando [uma série de [novas espécies [68]] de [um tipo [68]] todo especial de [fungo [157]] [169]] : [pequenos cogumelos [157]] que emitem [uma misteriosa luminosidade verde [51]] em [o escuro [74]] . As criaturas , antes desconhecidas em [o Brasil [72]] , podem ajudar a elucidar [o mecanismo bioquímico [167]] que leva a [a produção [136]] de [luz [141]] em [fungos [157]] . Além_disso isso , com um pouco mais de [estudo [32]] , poderiam servir como [sensores vivos de [poluição [140]] ou mesmo fontes de moléculas úteis para [a biotecnologia [32]] [139]] . Segundo [Cassius_Vinicius_Stevani [190]] , [químico [77]] de [a USP [29]] que coordena [os estudos [32]] , é possível [que o material recolhido [32]] abranja por o menos_dez [espécies novas [68]] . Não é pouca coisa , já que em o mundo todo se conhecem só 42 espécies de [o fungo [157]] , quase todas restritas a o Sudeste_Asiático . " Já temos

Fig. 2. Correference chains indicated by indexes and colours

Figures 2 and 3 show the output generated by the system. It generates both a coloured version of the text and a corresponding table containining all coreference chains. Note that there are some embedded mentions, such as “USP” in “pesquisadores da USP”. In those cases we present the larger expression in the same colour and use a different collor only for the brackets and ID of the inside mention. In the table, unique mentions (mentions that appear only once in the text) are also listed. We use the same colors in the text and table to represent the coreference chains. In this example we see that a semantic rule has been used when matching fungos and cogumelos (*funghi and mushroom*).

4 Conclusion

In this paper, we presented a rule-based coreference resolution system for Portuguese. We believe that this tool may help many researchers, due to fact that the coreference resolution task may help in several NLP tasks.As further work, we intend to enrich our semantic rules and develop other modules, such as pronominal coreference resolution. The CORP implementation is part of a PhD thesis entitled: “Resolução de Correferências em Língua Portuguesa” (*Coreference Resolution in Portuguese Language*).



	Tokens	Sintagma
CADEIA_68		
SnID: 4	12 ... 13	**novas espécies**
SnID: 5	15 ... 16	**um tipo**
SnID: 33	109 ... 110	**espécies novas**
SnID: 68	237 ... 238	**o tipo**
SnID: 71	251 ... 251	**espécies**
CADEIA_157		
SnID: 7	20 ... 20	**fungo**
SnID: 8	22 ... 23	**pequenos cogumelos**
SnID: 18	58 ... 58	**fungos**
SnID: 39	129 ... 130	**o fungo**
MençõesÚnicas:		
ID: 6	17 ... 18	**todo especial**
ID: 12	34 ... 36	**As criaturas**

Fig. 3. Coreference chains and unique mentions

Acknowledgments

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

References

1. S. Collovini, T. I. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, and R. Vieira. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana*, 2007.
2. W. D. C. de Moura Silva. Aprimorando o corretor gramatical cogroo. Master's thesis, Universidade de São Paulo, 2013.
3. E. B. Fonseca, R. Vieira, and A. Vanin. Adapting an entity centric model for portuguese coreference resolution. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*, In Press, 2016.
4. E. B. Fonseca, R. Vieira, and A. Vanin. Improving coreference resolution with semantic knowledge. In *Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016)*, In Press, 2016.
5. H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013.
6. H. G. Oliveira and P. Gomes. Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. In *Proceedings of Language Resources and Evaluation Conference*, volume 48, pages 373–393. Springer, 2014.