

Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts

Evandro Fonseca¹, Vinicius Sesti¹, Sandra Collovini¹, Renata Vieira¹,
Ana Luísa Leal² and Paulo Quaresma³

evandro.fonseca@acad.pucrs.br, vinicius.sesti@acad.pucrs.br,
sandra.abreu@acad.pucrs.br, renata.vieira@pucrs.br, analeal@umac.mo,
pq@di.uevora.pt

¹Pontifical Catholic University of Rio Grande do Sul (PUCRS)

²University of Macau

³University of Évora

Abstract. This paper describes the collaborative creation of a corpus with coreference annotation for Portuguese. The annotation was performed using the coreference annotation CORP, and the editing tool CorrefVisual. The texts were automatically annotated and manually revised by Portuguese speakers. As a result a new corpus for coreference studies was produced for Portuguese.

1 Introduction

In this paper we describe the creation of a collaborative annotated corpus. Seven teams participated in the task. The texts were chosen by the teams themselves. As a result of this task, we created ‘Corref-PT’, a coreference corpus for Portuguese. The texts submitted by the teams were first annotated with CORP [10], a nominal coreference resolution tool for Portuguese. Then, the editing tool CorrefVisual [28] was used for the manual revision of the previously annotated texts. Agreement was measured with Kappa, considering the concordance among team members and across teams.

The paper is organized as follows. Section 2 presents an overview of the problem of coreference resolution. Related work is presented in Section 3. Section 4 presents an overview of the corpus submission and information about participating teams. Section 5 describes the corpus annotation, including the distribution of texts among annotators, annotation tools and annotation agreement. Section 6 describes the results of this IberEval task: the Corref-PT corpus, its metrics and a brief discussion regarding the annotation process and its problems. Finally, in Section 7, the conclusions and future work are presented.

2 Coreference Resolution

Coreference resolution basically consists of finding different references to a same entity in a text, as in the example: “A França resiste como único país da União

29. O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, K. Rodriguez, and M. Poesio. ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2058–2062, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).
30. K. van Deemter and R. Kibble. What is coreference, and what should coreference annotation be? In *Proceedings of the Workshop on Coreference and Its Applications, CorefApp '99*, pages 90–96, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.
31. A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.