CRT-ML - Coreference Resolution Tool through Machine Learning

Jefferson F. Silva and João Luís G. Rosa

Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade da São
Paulo (USP), Av. Trabalhador São-Carlense, 400 – Centro, Caixa Postal: 668 - CEP:
13560-970 - São Carlos – SP
Fone: 55 (16) 3373-9700 - Fax: 55 (16) 3371-2238
jfs@icmc.usp.br, joaoluis@icmc.usp.br

Abstract. This paper presents the tool CRT-ML for annotation of coreference in Portuguese texts. The tool is based on a supervised machine learning technique that uses shallow features obtainable from the output of a part-of-speech tagger. For training and evaluation a corpus of Portuguese language with annotation of coreference was employed. CRT-ML adopts the XML format employed by MMAX tool in order to display the achieved annotation result.

1 Introduction

Tools of coreference resolution are useful for applications of automatic summarization, automatic translation, information extraction and questions answering system. It is difficult to find coreference resolution tools for Portuguese language and, when they exist, their integration with other systems is often hard to deal with.

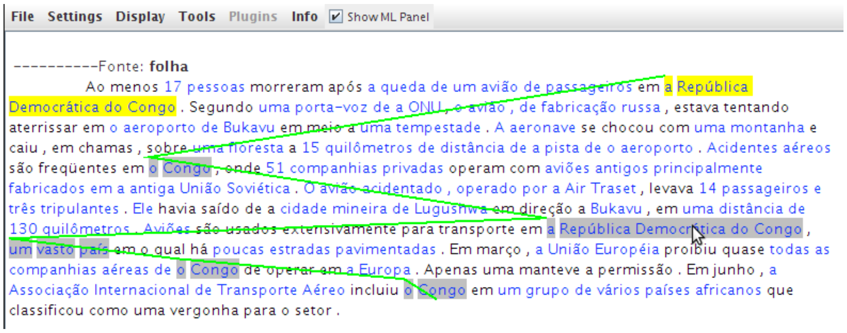
The proposed tool CRT-ML uses machine learning methods to identify the chains of coreference. This approach employs a corpus with annotation of coreference to train a classifier, and it was already succesfully employed [4].

2 Description

The method used to resolve the coreference implemented in CRT-ML is based on the work of Souza et al. [4]. This method uses a set of morphosyntactic and semantic features to obtain the chains of coreference. These features are extracted automatically using the morphosyntactic parser PALAVRAS [1]. This parser is used to recognize named entity as well as to obtain morphosyntactic features.

The tool annotates noun phrases and pronouns. In order to treat these two types of entities, the set of attributes is chosen appropriately so that it is possible to resolve the coreference for the mentioned entities. WEKA [2] is employed to train the classifier, through a decision tree J4.5, an implementation of C4.5 in WEKA.

The tool CRT-ML uses the XML format to represent the annotation of chains of coreference found in the text. This format is compatible with MMAX corpus annotation tool [3]. This allows the use of the tool in a automatic system or in a corpus annotation process. It is also available the visualization of the results provided by the tool, shown in Figure 1, where the chains of coreference of an entity are displayed. In this case, the tool can be used for an initial annotation which is then subject to human verification and correction.



----------Fonte: **folha**

Ao menos 17 pessoas morreram após a queda de um avião de passageiros em a República Democrática do Congo . Segundo uma porta-voz de a ONU , o avião , de fabricação russa , estava tentando aterrissar em o aeroporto de Bukavu em meio a uma tempestade . A aeronave se chocou com uma montanha e caiu , em chamas , sobre uma floresta a 15 quilômetros de distância de a pista de o aeroporto . Acidentes aéreos são freqüentes em o Congo , onde 51 companhias privadas operam com aviões antigos principalmente fabricados em a antiga União Soviética . O avião acidentado , operado por a Air Traset , levava 14 passageiros e três tripulantes . Ele havia saído de a cidade mineira de Lugushwa em direção a Bukavu , em uma distância de 130 quilômetros . Aviões são usados extensivamente para transporte em a República Democrática do Congo , um vasto país em o qual há poucas estradas pavimentadas . Em março , a União Européia proibiu quase todas as companhias aéreas de o Congo de operar em a Europa . Apenas uma manteve a permissão . Em junho , a Associação Internacional de Transporte Aéreo incluiu o Congo em um grupo de vários países africanos que classificou como uma vergonha para o setor .

Fig. 1. Example of annotation held by CRT-ML

3 Final Remarks

The tool CRT-ML was implemented using Groovy and Java languages. CRT-ML was designed to facilitate expansion of features and also substitution of existing ones. It is available under the GNU-PL v3.0 license.

Acknowledgements. We thank CAPES for the master's thesis financial support. CRT-ML is being developed as part of this thesis.

References

1. Eckhard Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press, 2000.
2. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, *The WEKA Data Mining Software: An Update*, SIGKDD Explorations **11** (2009), no. 1, 10–18.
3. Christoph Mueller and Michael Strube, *MMAX: A Tool for the Annotation of Multimodal Corpora*, Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2001, pp. 45–50.
4. José Guilherme Souza, Patricia Nunes Gonçalves, and Renata Vieira, *Learning Coreference Resolution for Portuguese Texts*, Proceedings of the 8th international conference on Computational Processing of the Portuguese Language(Lecture Notes In Artificial Intelligence; Vol. 5190) (Berlin, Heidelberg), Springer-Verlag, 2008, pp. 153–162.