An Analysis of Sentence Boundary Detection Systems for English and Portuguese Documents*

Carlos N. Silla Jr. and Celso A. A. Kaestner

Pontifical Catholic University of Parana
Rua Imaculada Conceicao, 1155 - 80.215-901
Curitiba - Parana - BRAZIL
{silla, kaestner}@ppgia.pucpr.br

Abstract. In this paper we present a study comparing the performance of different systems found in the literature that perform the task of automatic text segmentation in sentences for English documents. We also show the difficulties found to adapt these systems to make them work with Portuguese documents and the results obtained after the adaptation. We analyzed two systems that use a machine learning approach: MxTerminator and Satz, and a customized system based on fixed rules expressed by Regular Expressions. The results achieved by the Satz system were surprisingly positive for Portuguese documents.

1 Introduction

When dealing with tasks related to the automatic processing of documents like summarization, translation, etc. one of the procedures that frequently occur is the segmentation of the text in sentences. This task is usually included in the pre-processing stage, and uses a simple criterion, tagged documents, or one of the approaches found in the literature.

The systems found in the literature can be grouped in two classes: the ones that use fixed rules to identify what is and what is not a sentence, and the ones that use a machine learning approach. In this work we evaluate the performance of one customized system that uses fixed rules, and two systems that use a machine learning approach: MxTerminator [1] and Satz [2]. The first system uses templates based on Regular Expressions, considering the context where a punctuation mark appears, and will be refered to as RE (Regular Expressions) [3]. The MxTerminator uses a Maximum Entropy Model to detect the sentence boundaries, while Satz considers the context where a possible punctuation mark appears and can be used with any machine learning algorithm; in this work, it was used with the C4.5 classifier [4].

The remaining part of the article is divided as follows: section 2 presents a general view of the systems used for comparison and how they were adapted to Brazilian Portuguese; section 3 describes the methodology used in the experiments and presents the corresponding results for two sets of documents, in

* This research was supported by the Brazilian PIBIC-CNPq Agency.