A Comparative Evaluation of a New Unsupervised Sentence Boundary Detection Approach on Documents in English and Portuguese

Jan Strunk[1], Carlos N. Silla Jr.[2], and Celso A. A. Kaestner[2]

[1] Sprachwissenschaftliches Institut, Ruhr-Universität Bochum,
44780 Bochum, Germany
strunk@linguistics.rub.de
[2] Pontifical Catholic University of Paraná,
Rua Imaculada Conceição 1155, 80215-901 Curitiba, Brazil
{silla,kaestner}@ppgia.pucpr.br

Abstract. In this paper, we describe a new unsupervised sentence boundary detection system and present a comparative study evaluating its performance against different systems found in the literature that have been used to perform the task of automatic text segmentation into sentences for English and Portuguese documents. The results achieved by this new approach were as good as those of the previous systems, especially considering that the method does not require any additional training resources.

1 Introduction

We are living today in an era of information overload. The web alone contains about 170 terabytes of information, which is roughly 17 times the size of the printed material in the Library of Congress of the USA; cf. [1]. However, it is becoming more and more difficult to use the available information. Many problems such as the retrieval and extraction of information and the automatic summarization of texts have become important research topics in computer science. The use of automatic tools for the treatment of information has become essential to the user because without those tools it is virtually impossible to exploit all the relevant information available on the Web.

One pre-processing component that is essential to most text-based systems is the automatic segmentation of a text into sentences. Existing systems for sentence boundary detection mostly either use a set of heuristics or a supervised machine learning approach. The drawback of both these approaches is that adapting them to new languages can be time and resource intensive. In the first case, it is necessary to adapt the rules to the new language. In the second case, a new training corpus has to be tagged manually for retraining.

In this paper, we compare a new unsupervised approach to sentence boundary detection by Kiss & Strunk [2] with the results of a previous evaluation of three