A Paragraph Boundary Detection System

Dmitriy Genzel

Department of Computer Science
Box 1910
Brown University
Providence, RI 02912, USA
dg@cs.brown.edu

Abstract. We propose and motivate a novel task: paragraph segmentation. We discuss and compare this task with text segmentation and discourse parsing. We present a system that performs the task with high accuracy. A variety of features is proposed and examined in detail. The best models turn out to include lexical, coherence, and structural features.

1 Introduction

In this paper we will introduce the problem of paragraph boundary detection (or paragraph segmentation). Given a collection of semantically coherent texts, such as news articles or book chapters, with sentence boundaries marked, we wish to mark the paragraph boundaries in each text.

The system that solves this problem has several applications. The most interesting one is to be used as a part of a grammar checker in a word processing system like Microsoft Word [1]. The system would suggest places where a paragraph could be split in two, or even suggest a paragraph break as soon as the user typed in an appropriate sentence.

Furthermore, it could be used to restore paragraph breaks after OCR processing in the cases when the OCR module is unable to accomplish this based on spatial information. This happen fairly often when the paragraph starts at the top of the page.

It is also the case that some text summarization systems prefer extracting paragraphs from original text [2], and thus can use this system when the paragraph boundaries are not specified in the text.

This problem has received very little attention, although it is similar to several well-known tasks, such as text segmentation and discourse parsing. The problem is interesting in its own right as well as through its potential applications. By building such a system, one can obtain information about the regularity of paragraph boundary location. After all, the paragraph boundary placement is somewhat arbitrary and depends on the author's style and taste. Furthermore, by examining the most relevant features of the model, one can learn which elements of the text are sensitive to paragraph boundary placement. This knowledge could then be used to better predict these features based on the paragraph boundary information, which has been rarely used in models, even when available.