Creating a Testbed for the Evaluation of Automatically Generated Back-of-the-book Indexes

Andras Csomai and Rada Mihalcea

University of North Texas
Computer Science Department
csomaia@unt.edu, rada@cs.unt.edu

Abstract. The automatic generation of back-of-the book indexes seems to be out of sight of the Information Retrieval and Natural Language Processing communities, although the increasingly large number of books available in electronic format, as well as recent advances in keyphrase extraction, should motivate an increased interest in this topic. In this paper, we describe the background relevant to the process of creating back-of-the-book indexes, namely (1) a short overview of the origin and structure of back-of-the-book indexes, and (2) the correspondence that can be established between techniques for automatic index construction and keyphrase extraction. Since the development of any automatic system requires in the first place an evaluation testbed, we describe our work in building a gold standard collection of books and indexes, and we present several metrics that can be used for the evaluation of automatically generated indexes against the gold standard. Finally, we investigate the properties of the gold standard index, such as index size, length of index entries, and upper bounds on coverage as indicated by the presence of index entries in the document.

1 Introduction

> *"Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information on it." (Samuel Johnson)*

The automatic construction of back-of-the-book indexes is one of the few tasks related to publishing that still requires extensive human labor. While there is a certain degree of computer assistance, mainly consisting of tools that help the professional indexer organize and edit the index, there are however no methods or tools that would allow for a complete or nearly-complete automation. Despite the lack of automation in this task, there is however another closely related natural language processing task – kepyphrase extraction – where in recent years we have witnessed considerable improvements.

In this paper, we argue that the task of automatic index construction should be reconsidered in the light of the progress made in the task of keyphrase extraction. We show how, following methodologies used for the evaluation of keyphrase extraction systems, we can devise an evaluation methodology for back-of-the-book indexes, including a gold standard dataset and a set of evaluation metrics.