Challenges in the Interaction of Information Retrieval and Natural Language Processing *

Ricardo Baeza-Yates

Center for Web Research
Dept. of Computer Science, University of Chile
Blanco Encalada 2120, Santiago, Chile
E-mail: rbaeza@dcc.uchile.cl

Abstract. In this paper we explore the challenges to effectively use natural language processing (NLP) for information retrieval. First, we briefly cover current NLP uses and research areas in the intersection of both fields, namely summarization, information extraction, and question answering. Second, we motivate other possible challenging uses of NLP for information retrieval such as determining context, semantic search, and supporting the Semantic Web. We end with a particular use of NLP for a new problem, searching the future, that poses additional NLP challenges.

1 Introduction

The interaction of Information Retrieval (IR) and Natural Language Processing (NLP) has two sides. The focus of this paper is NLP for IR. On the other hand, IR has always been a useful tool for NLP and will continue to be so. A short account of NLP research is given by Sparck-Jones in [43], but she also analyzes the accomplishments until 1994 and proposes summarization as the next step ahead [43]. The state of the art in natural language processing is covered in [6, 11, 18, 31].

The interaction, the role, the evaluation, and the progress of NLP for IR is covered in [44, 42, 45, 37], respectively. Regarding the evaluation of the results, several authors point out that the improvements of using sophisticated NLP techniques are too small to justify their cost compared with statistical IR techniques [51, 24, 41]. Even the use of NLP resources such as thesauri coupled with IR techniques was discouraging [41]. Two issues are described and analyzed in [24]: (1) whether more refined natural language indexing is wanted, and, (2) whether controlled language indexing is really needed. Both imply non-trivial NLP research, such as indexing multiword expressions and finding semantic relations.

* Partially funded by Fondecyt Grant 1020803 of CONICYT Chile. This paper was written while visiting the Dept. of Computer Science and Software Engineering of the University of Melbourne, Australia.