Head/Modifier Frames for Information Retrieval

C.H.A. Koster

Computing Science Institute,
University of Nijmegen,
The Netherlands,
E-mail: kees@cs.kun.nl

Abstract. We describe a principled method for representing documents by phrases abstracted into Head/Modifier pairs. First the notion of aboutness and the characterization of full-text documents by HM pairs is didcussed. Based on linguistic arguments, a taxonomy of HM pairs is derived. We briefly describe the EP4IR parser/transducer of English and present some statistics of the distribution of HM pairs in newspaper text. Based on the HM pairs generated, a new technique to measure the accuracy of a parser is introduced, and applied to the EP4IR grammar of English. Finally we discuss the merits of HM pairs and HM trees as a document representation.

1 Introduction

The Information Retrieval community has for a long time held high hopes concerning the value of linguistic techniques. However, the improvements in precision and/or recall expected from the use of phrases in retrieval and in text categorization have repeatedly been found disappointing [22].

Although the use of simple noun phrases as indexing terms is now commonly accepted, practical Information Retrieval systems using phrases like the CLARIT system [7] do not appear to perform consistently better than those based on keywords. There is a growing conviction that the value of Natural Language Processing to IR is dubious, even among people who tried hard to make linguistically-based IR work [15, 20]. The predominant feeling, as voiced in [18], is that only 'shallow' linguistic techniques like the use of stop lists and lemmatization are of any use to IR, the rest is a question of using the right statistical techniques.

In spite of these negative experiences, we are trying to improve the accuracy of automatic document classification techniques by using (abstractions of) phrases as terms. In this paper we shall first discuss the notion of *aboutness*, which plays a central role in Information Retrieval. We then introduce Head/Modifier (HM) pairs as an abstraction of phrases preserving their aboutness, and give a taxonomy of HM pairs based on the intra-sentence relations they represent. We describe the EP4IR grammar, in which the transduction of English text to HM pairs is realized, and which is now available in the public domain.