Merging Case Relations into VSM to Improve Information Retrieval Precision

Wang Hongtao,[1] Sun Maosong,[1] Liu Shaoming[2]

[1] The State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
wanght02@mails.tsinghua.edu.cn
[2] Future Technology Institute, Fuji Xerox Co. Ltd, Japan
Liu.shaoming@fujixerox.co.jp

Abstract. This paper presents an approach that merges case relations into the well-known Vector Space Model (VSM), leading to a new model named C-VSM (Case relation-based VSM). A Chinese case system with 23 case relations is established, and a Chinese Olympic news corpus of 7,662 sentences, denoted COCS, is constructed by manual annotation with these 23 case relations. We use 50 queries on COCS as a test set. Experimental results on the test set show that C-VSM outperforms W-VSM (Word-based VSM) by 3.4% on the average 11-point precision. It is worth pointing out that almost all the previous studies on semantic IR obtained no better, even worse, results than W-VSM, our work thus validates the usefulness of case relations in IR through the validation is still preliminary. The proposed model is believed to be language-independent.

1 Introduction

A majority of traditional models of information retrieval (IR) mainly make use of surface linguistic information such as words/terms. It is reasonable to expect better retrieval results if we can exploit deep linguistic information further. Previous studies of this sort have been carried out at both syntactic and semantic levels. Most of them focused on the former, because the recognition of syntactic structures is easier than that of semantic structures. Syntactic information possibly exploited in IR can be a simple syntactic relation between a pair of words, and can also be a complex structure tree. The use of simple syntactic relations in IR has found a small improvement in retrieval effectiveness (Croft, Turtle and Lewis, 1991; Hyoudo, Niimi and Ikeda, 1998). But the results of using complex structure trees are worse than keyword matching (Smeaton, O'Donnell and Kelledy, 1995).

It is natural to assume that semantic information is more useful in IR since it can capture the meaning of a sentence more precisely than syntax. Semantic information, both intra-sentential and inter-sentential, is usually represented by the so-called semantic relations between various entities involved.

Case relation is an intra-sentential semantic relation that exists between the core verb and other constituents of a sentence (Fillmore, 1968; Somers, 1987). Lewis (1984) addressed the possibility of IR based on case relation matching. Lewis' major