电子与信息学报 2008 30 (3): 725-729 ISSN: 1009-5896 CN: 11-4494/TN

论文

基于统计机器翻译模型的查询扩展

李卫疆, 赵铁军, 王宪刚

哈尔滨工业大学计算机科学与技术学院语音语言教育部-微软重点实验室 哈尔滨 150001

收稿日期 2006-9-26 修回日期 2007-1-26 网络版发布日期 2008-6-5 接受日期

摘要
在搜索引擎等实际的信息检索应用中,用户提交的查询请求通常都只包含很少的几个关键词,这会引起相关文档与用户查询之间的词不匹配问题,对检索性能有较严重的负面影响。该文在分析了查询产生模型的基础上,提出了一种新的基于统计机器翻译的查询扩展方法。通过统计机器翻译模型提取文档集中与查询词相关联的词,用以进行查询扩展。在TREC数据集上的试验结果表明:基于统计翻译的查询扩展方法不仅比不扩展的语言模型方法始终有12%～17%的提高,而且比流行的查询扩展方法-伪反馈也具有可比的平均准确率。

关键词 信息检索 查询扩展 语言模型 统计机器翻译

分类号 TP391

A SMT-based Approach for Query Expansion in Information Retrieval

Li Wei-jiang, Zhao Tie-jun, Wang Xian-gang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, china

Abstract
In practical applications of information retrieval, such as the search engine，the query user submitted contains only several keywords usually. This will cause unmatched issue of word of relevant files and user’s query and have more serious negative effects on the performance of information retrieval. On the basis of analyzing of process of producing query, this paper puts forward a new method of query expansion on the basis of model of statistical machine translation. The approach extract related terms between documents and query through statistical machine translation model, then expand into query. The experiment result on TREC data collection shows the proposed method, SMT-based query expansion, has 12 - 17% of the improvement all the time more than the language model method without expanding. Compared to the popular approach of query expansion, pseudo feedback, the proposed method has the competed average precision.

Key words Information retrieval Query expansion Language model Statistical Machine Translation（SMT）

DOI:

通讯作者

作者个人主页 李卫疆; 赵铁军; 王宪刚