第 33 卷 第 11 期 计 算 机 工 程 2007 年 6 月
Vol.33 No.11 Computer Engineering June 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)11—0166—02 文献标识码：A 中图分类号：TP391

跨语言信息检索中查询语句翻译转换算法

张孝飞，黄河燕，陈肇雄，代六玲
（中国科学院计算机语言信息工程研究中心，北京 100083）

摘 要：跨语言信息检索中，输入的查询语句往往是一系列关键词组合，而不是一个完整意义上的句子，致使查询关键词序列缺乏必要的语法、语境信息，难以实现查询语句的精确翻译。该文基于大规模双语语料库，以向量空间模型和词汇同现互信息为理论基础，运用传统单语信息检索技术，将查询语句的翻译问题转换为查询关键词词典义项的 boost 值计算，重构目标语查询语句。

关键词：跨语言信息检索；查询语句；翻译转换；双语语料库

Translation Transforming Algorithm of Query Sentence in Cross Language Information Retrieval

ZHANG Xiaofei, HUANG Heyan, CHEN Zhaoxiong, DAI Liuling
(Research Center of Computer & Language Information Engineering, Chinese Academy of Sciences, Beijing 100083)

【Abstract】In cross language information retrieval, the query sentence often comprises some query keywords, but not a complete sentence. Because of the lack of necessary contextual and syntactic information in the series of query keywords, it is impossible to accurately translat the query sentence. In this paper, based on large-scale bilingual corpora and theory of vector space model and lexical mutual information, traditional mono-language IR technology is applied to convert the problem of translating query sentence into computing the “boost” value of the query keyword translation in bilingual dictionary, and the target language query sentence is reconstructed.

【Key words】 Cross language information retrieval; Query sentence; Translation transforing; Bilingual corpora

信息检索(information retrieval，IR)泛指用户从包含各种信息的文档集中找到所需要的信息或知识的过程。传统的信息检索系统主要针对单一语种的文档集实现，一般使用用户最为熟悉的语种作为查询语言。而随着互联网的迅速发展，用户面对查询一个多语种文本集合的情形变得越来越普遍[1]。这就产生了一个难题——以一种语言描述的用户查询与其他不同语种书写的文本之间的匹配问题，也就是如何跨越语言界限的问题[2,3]，即跨语言信息检索（cross language information retrieval，CLIR）。

跨语言信息检索是指用某一种语言提出检索要求，计算机在其他不同语种的文本中进行自动搜索，得到的检索结果甚至可以翻译成用户指定的特定语种。跨语言信息检索结合了传统文本信息检索技术和机器翻译（machine translation，MT）技术[4,5]。在当今信息社会中，跨语言信息检索已成为世界范围内一个亟待解决的关键课题。事实上，在TREC中就多次安排设计了跨语言信息检索任务。

1 问题描述和转换

在跨语言信息检索系统中，用户用某一种语言提出检索要求，计算机在其他不同语种的文本中进行自动搜索。因而必须解决如何跨越语言界限的问题。大多数跨语言信息检索系统采用 MT 的查询翻译作为跨越源语与目标语之间语言界限。该方法通常以双语词典作为主体知识源完成查询翻译处理过程。

MT 的输入通常是一句完整的句子。完整的句子一般具有比较充分的上下文信息和语法语义信息，这样经过 MT 的翻译才比较准确。但是，在信息检索中，通常输入的并不是一个完整的句子，而是一系列关键词组合。这些查询关键词序列由于缺乏必要的语境、语法信息，所以不能简单地采用传统的 MT 技术进行翻译。另外，也不能简单地通过查询双语词典解决这种翻译问题。如“bank”词典义项中有“(1)银行；(2)堤，岸”等，那么 bank 到底应该翻译成银行还是堤岸呢？

根据常识，在用户查询“bank; credit”中，很可能是查询有关“银行、信用”方面的信息，但也不能排除“bank”在这里释义为“河岸”的可能，用户也许就是想查询“是否有人在河岸边捡到了他的信用卡”，因而在该查询中，“bank”既可能是“银行”的意思，也可能是“河岸”的意思，只不过翻译成“银行”的可能性比翻译成“河岸”的可能性大得多，即算法应该使得检索结果中包含“银行”这个关键词的文档排在较前位置，而包含“河岸”这个关键词的文档排在较后位置。

跨语言信息检索中查询语句翻译转换任务的形式化描述如下：

设有源语言关键词查询语句：

$$Query_S = W_{S1} W_{S2} \ldots W_{Si} \ldots W_{SN} \quad (1)$$

其中，W_{Si}指查询语句中第 i 个查询关键词，比如查询实例“bank credit”中的“bank”和“credit”。

基金项目：国家自然科学基金资助项目(60502048)
作者简介：张孝飞(1970－)，男，副研究员、博士，主研方向：自然语言处理，信息检索，机器翻译；黄河燕、陈肇雄，研究员、博导；代六玲，讲师、博士
收稿日期：2006-06-07 **E-mail**：zxflying@gmail.com

翻译转换后得到新的目标语关键词查询语句，表示为

$$Query_T = (W_{T11} \wedge boost_{T11} \quad W_{T12} \wedge boost_{T12} \cdots W_{T1N} \wedge boost_{T1N}) \\ \ldots (W_{Ti1} \wedge boost_{Ti1} W_{Ti2} \wedge boost_{Ti2} \cdots W_{TiN} \wedge boost_{TiN}) \ldots \\ (W_{TN1} \wedge boost_{TN1} W_{TN2} \wedge boost_{TN2} \cdots W_{TNN} \wedge boost_{TNN}) \quad (2)$$

其中，$W_{T11} \quad W_{Ti2} \cdots W_{TiN}$ 是源语言关键词查询语句中关键词 W_{Si} 在双语词典中的词典义项；$boost_{Ti1}\ boost_{Ti2} \ldots boost_{TiN}$ 是关键词 W_{Si} 的各个词典义项的权值，称之为 $boost$ 值。

跨语言信息检索的关键问题之一是如何跨越语言界限，即如何由源语言关键词查询语句形成目标语关键词查询语句。至此把查询语句的翻译问题转换为词典义项的 $boost$ 值计算问题。依据一定的策略计算词典义项的 $boost$ 值即为本算法的关键。

2 算法分析

2.1 信息检索的基本理论

对于一个检索系统，核心是排序问题，即所谓的相关性计算。基于向量空间模型设计了如下相关度计算公式：

$$score(q,d_i) = \frac{\sum_{t \in q} tf(t \text{ in } d_i) idf(t) boost(t)}{\max(score(q,d_i))} \quad (3)$$

其中，q 表示查询语句，d_i 表示第 i 篇文档，t 表示查询关键词。分母 $\max(score(q,d_i))$ 是归一化因子，不影响排序结果，但是使得不同查询的相关度分值有一定的可比性。此外，加入归一化因子也是为了方便查询关键词翻译转换时 $boost$ 值的计算。式(3)中 $idf(t)$ 的计算公式如下：

$$idf(t) = -\log p(t) = -\log(\frac{N}{M}) \quad (4)$$

其中，M 表示文档总数，N 表示包含词汇 t 的文档数。

2.2 $boost$ 值计算方法

查询关键词词典义项的 $boost$ 值计算，是本算法的重点。本文基于大规模双语语料库，以向量空间模型和词汇同现互信息为理论基础，运用传统单语信息检索技术，实现了一种查询关键词词典义项的 $boost$ 值计算方法。

(1)支撑知识库

查询语句翻译转换所用到的知识库，除了双语词典外，还用到了一个大规模双语句对齐语料库，即以双语句对作为最基本的检索单元。实验中使用的语料库总共包含 162、918 对英汉双语句对，大约 200 万词的英文和 200 万词的中文。其中英文平均句长为 12.5 个词左右，中文平均句长为 11 个词左右。这个双语句对齐语料库作为计算 $boost$ 值时的查询检索源。

(2) $boost$ 值计算

设有源语言查询关键词序列 $W_{S1}\ W_{S2} \cdots W_{Si} \cdots W_{SN}$ 以及对应的词典义项：

$$(W_{T11}\ W_{T12} \ldots W_{T1N}) \ldots (W_{Ti1}\ W_{Ti2} \ldots W_{TiN}) \ldots (W_{TN1}\ W_{TN2} \ldots W_{TNN})$$

为了计算查询关键词词典义项的 $boost$ 值，设计了 3 种查询语句：

$$Query_1 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ AND\ W_{S2} \ldots AND\ W_{SN}) \quad (5)$$

$$Query_2 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ OR\ W_{S2} \ldots OR\ W_{SN}) \quad (6)$$

$$Query_3 = (W_{Si}\ AND\ W_{Tij})\ OR\ (W_{S1}\ OR\ W_{S2} \ldots OR\ W_{SN}) \quad (7)$$

则词典义项 W_{Tij} 的 $boost$ 值计算公式如下：

$$boost(W_{Tij}) = 2^{\alpha} \times mean(score(q,d_i)) + \beta \quad (8)$$

其中，$mean(score(q,d_i))$ 表示查询结果的平均相关度，实际中为了提高处理速度只计算了前 100 篇文档的平均相关度，相关度计算公式参见式(3)；β 相当于转换基准值，设为 0.5；α 是查询语句的权值系数，按以下方式确定：

$$\alpha = \begin{cases} 3, & if\ q == Query_1 \\ 2, & else\ if\ q == Query_2 \\ 1, & else\ if\ q == Query_3 \\ 0, & \text{其他} \end{cases} \quad (9)$$

3 实验设计和结果

3.1 测试语料

从互联网上随机选取了一些中文网站进行下载，包括新华网、搜狐、新浪等。文件总共约 7GB，包括 138 908 个网页。这些网页全部作为测试语料。

3.2 测试方案

实验的目的是测试本文查询语句翻译转换算法的有效性，实验步骤如下：

第 1 步 精心设计了 100 组英文关键词查询语句。每一组查询语句中，至少有一个查询关键词翻译成中文时其语义有比较明显的差别。比如下面的一组关键词查询：

1){bank；credit}，“ bank ”翻译成中文可以是“ 银行；河岸 ”等，由于此例中很可能是“ 银行 ”的意思，因此检索结果中，包含“ 银行 ”这个关键词的文档应该排在较前位置。

2){bank；willow}，“ bank ”在这个查询语句中很可能是“ 河岸 ”的意思，在故检索结果中，包含“ 河岸 ”这个关键词的文档应该排在较前位置。

第 2 步 将这 100 组关键词查询语句提交系统进行英→中跨语言检索。

第 3 步 对检索结果进行人工判断。实验中只判断检索结果中前 10 个文档和前 100 个文档的正确性。评判标准是：如果文档中包含了正确翻译的词汇，则认为该文档是正确的查询结果，否则认为是错误的查询结果。比如：查询{bank；credit}，如果查询结果中的某篇文档包含“ 银行 ”这个关键词，就认为该文档是正确的查询结果；如果包含“ 堤，岸 ”等关键词，则认为该文档是错误的查询结果。

第 4 步 作为对比，查询关键词翻译转换时不采用本算法，而是直接查找英汉双语词典（词典方法），重复第 2 步和第 3 步。

3.3 实验结果与分析

实验结果如表 1 所示。在 100 个查询的前 10 个文档中，正确文档数达 928 篇，正确率为 92.8%，比词典方法的正确率高出 15.4%多。在 100 个查询的前 100 个文档中，正确文档数也达到 8 891 篇，正确率为 88.9%，比词典方法的正确率高出 13.0%。结果表明，本文提出的查询关键词翻译转换算法较为有效。

表 1 对比试验结果

算法	前 10 篇正确总数	前 10 篇正确率	前 100 篇正确总数	前 100 篇正确率
本文方法	928	92.8%	8 891	88.9%
词典方法	774	77.4%	7 593	75.9%

此外，还分析了前 100 篇正确率偏低的原因（正确率只有 88.9%）。主要原因之一是测试语料库规模较小，使得某些查询结果的文档总数偏少，比如{bank；willow}的查询结果文档总数只有 260 篇，影响了整个的查询准确率。

4 结论和进一步的研究

本文基于大规模双语语料库，以向量空间模型和词汇同现互信息为理论基础[6,7]，运用传统单语信息检索技术，将查询语句的翻译问题转换为关键词词典义项的 $boost$ 值计算，从而重构目标语查询语句。实验结果表明，本算法基本满足跨语言信息检索的实际要求。 （下转第 212 页）