面向英汉的跨语言信息检索关键技术研究*

张玥杰[1] 郭依昆[2] 吴立德[3]
复旦大学计算机科学与工程系，上海 (200433)
E-mail: yjzhang@fudan.edu.cn

摘 要：本文以参加文本检索会议中有关跨语言信息检索(Cross-Language Information Retrieval, CLIR)任务的评价作为研究背景，提出了一个面向英汉的 CLIR 系统的实现框架，并由此引出有关英汉机译方法及汉语信息检索策略的研究。其中，以查询翻译为主导策略，以英语查询作为翻译对象，并采取英汉双语词典作为获取翻译知识的重要知识源；同时，结合所构建的汉语 IR 系统，实现完整的英—汉 CLIR 过程。

关键词：信息检索，跨语言信息检索，机器翻译，语料库

Research on the Key Technique of English-Chinese-oriented Cross-Language Information Retrieval

Zhang Yuejie, Guo Yikun and Wu Lide
Department of Computer Science & Engineering
Fudan University, Shanghai (200433)
E-mail: yjzhang@fudan.edu.cn

Abstract: Taking the attendance in the Cross-Language Information Retrieval (CLIR) task evaluation of the Text Retrieval Conference (TREC) as the research background, we propose a kind of implementation frame of English-Chinese-oriented CLIR System. Based on this frame, we research the method of English-Chinese Machine Translation and the strategy of Chinese Information Retrieval. In this system, we adopt query translation as the dominant strategy, use English query as translation object, and utilize English-Chinese bilingual dictionary as the important knowledge resource to acquire correct translations. So combining the Chinese Information Retrieval System established by us, the complete English-Chinese CLIR process can be implemented successfully.

Key words: Information retrieval, cross-language information retrieval, machine translation, corpus.

---

*本文受国家自然科学基金（编号：60203010）和国家 863 基金（编号：2001AA114120）资助。

[1] 张玥杰，女，1973 年生，副教授，主要研究领域为机器翻译、中文信息处理及相关技术。

[2] 郭依昆，男，1973 年生，博士，主要研究领域为中文信息处理及相关技术。

[3] 吴立德，男，1937 年生，教授，博士生导师，主要研究领域为计算语言学、计算机图形学及相关技术。

1 前言

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是针对单一语种的文档集实现，一般是使用用户最为熟悉的语种作为查询语言。而随着日益增长的大量信息成为可利用的，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。这就产生一个非常重要的问题—以一种语言描述的用户查询与以不同语言书写的文本之间的匹配问题，也就是一种如何跨越语言界限的问题，即跨语言信息检索(Cross-Language Information Retrieval, CLIR)。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问题。

文本检索会议(Text Retrieval Conference, TREC)，是由美国国家技术标准局组织召开的国际会议，其旨在促进大规模文本检索领域的研究，加速研究成果向商业应用的转化，促进学术研究机构、商业团体和政府部门之间的交流与合作。跨语言信息检索是在第六届文本检索会议(TREC-6)评价中建立的一项新任务。在第九届文本检索会议(TREC-9)的 CLIR 任务评价中，第一次引入汉语作为文本描述语言。由此，我们的自然语言处理研究组参加了有关英-汉 CLIR 的任务评价，针对该项任务建立了一个面向英汉的跨语言信息检索系统，并以此为基础获取了相关的几组运行结果。其中，以查询翻译为主导策略，以英语查询作为翻译对象，并采取英汉双语词典作为获取翻译知识的重要知识源；同时结合所构建的汉语 IR 系统，实现完整的英—汉 CLIR 过程。

2 系统概述

在所建立的面向英汉的跨语言信息检索系统中，我们采取基于 MT 的查询翻译作为跨越源语与目标语之间所存在语言界限的方法，并利用以英汉双语词典为主体的知识源完成查询翻译处理过程。系统实现策略的基本思想是，首先，将初始的源语（英语）查询翻译为目标语（汉语）单词列表；然后，基于经过翻译处理的查询，利用汉语 IR 技术以及概率方法获取相关文档列表。其中，所有查询处理过程以完全自动的方式进行，同时涵盖长查询与短查询的翻译处理[4]。

2.1 基于英汉双语词典的查询翻译

2.1.1 知识源构建

在所建立的 CLIR 系统中，所涉及到的知识主要包括词典知识、汉语《同义词词林》、禁用词表以及单词形态恢复表。实际上，词典知识是一个知识表达和存储的载体，它包含词汇的几乎所有信息，这部分信息又称为静态信息。这四类知识通过合理的组织与有机的结合，

[4]查询按长度分为三类：标题查询（仅包含标题域中的单词）、短查询（包含描述域中的单词）以及长查询（包含主题描述中的所有单词）。

形成一个完整良好的机器翻译知识体系，描述如下：

(1) 英汉双语词典

主要用于进行词汇层和短语层翻译。整个英汉双语词典包含约 60,000 个词法条目。

(2) 汉语《同义词词林》

该词典实际上是一本义类词典，其中收录词语近 70,000 条，全部按所规范的语义关系进行编排，主要用于对经过翻译处理的翻译知识进行扩展，即查询扩展。

(3) 英语禁用词表

主要用于标注英语查询中的禁用词，包含 546 个英文常用词，如 is, one, exactly 等。

(4) 英语形态恢复表

实际上，该英语形态恢复表是一个词汇不规则变化表，包括名词、动词以及形容词的各种不规则变化形式，主要用于对具有不规则变化形式的单词进行形态恢复。它包含 12,555 个记录，其中每个记录反映单词原形及其不规则变化形式之间的关系。

2.1.2 预处理

英语查询预处理的主要任务是完成英语查询的分词、标点符号加标以及单词首写字母大小写变换处理三个过程，其各部分功能描述如下：

(1) 分词

对于英语查询，首先使用关于不同标点符号的启发式将其分割为句子。然后，利用空格作为标志，将一个句子的字符流切分为单词流，即完成分词过程。

(2) 标点符号加标

对于经过分词处理所获得的单词流，将其中的标点符号，如双引号[“”]、单引号[‘’]、冒号[:]、逗号[,]以及句末的句号[.]等，进行标注处理。需要注意的是，应该正确区分字符串所包含的[.]与句末的句号。

(3) 大小写变换

由于英语查询普遍为新闻报道的标题，首字母为大写形式的词汇较多，因此，必须对这种单词进行适当的大小写变换处理，以为后续的相关操作提供完整正确的信息。

2.1.3 预分析

基于经过预处理的英语查询，预分析主要完成三项任务，其一是对英语查询中的禁用词加以标注；其二是对其中具有变化形式的单词进行形态恢复处理；而其三即为对每一单词进行词类分析，确定其所属正确词性。

(1) 禁用词标记

考虑到翻译处理将涉及到某些禁用词，而有些禁用词无需翻译处理，因此利用禁用词表对禁用词进行标记，并结合翻译层次加以适当处理。

(2) 形态恢复

由于英语查询中存在一些具有变化形式的单词，这将不利于获取正确的翻译知识，因此必须利用英汉双语词典、针对不规则变化的形态恢复表以及规则变化的启发式对这些单词进行形态恢复处理，获取其相对应的原形。

(3) 词类分析

一个词可能含有多种不同的词性，在不同的句子中，承担不同的语法功能。因此，要决定一个词所应选择的词类，必须在具体句子中根据其它词的语法功能来互相判定。

词类分析问题过程的实现，是以我们自然语言处理研究组所建立的一种基于隐马尔可夫模型(Hidden Markov Mode, HMM)的词性标注器为基础，从而实现正确标注词类。

2.1.4 翻译处理

对于经过预处理与预分析处理的英语查询，其翻译过程主要分为两个层次进行：词汇层与短语层，最终获取与英语查询相对应的正确翻译知识。

(1) 词汇层翻译

该过程主要是利用英汉双语词典中的基本词典部分来实现逐词翻译，其中涉及到词义消歧的问题。一个单词可能对应多种不同的意义，词义是与具体词相联系的，脱离具体语境，词义也无法给出。而语境条件可以是语法语义参数，选取具体的词时就应该选中该词项的区别标记，这一标记应该代表一定语法和语义特征，并且唯一标识该词词义，称之为概念码。概念码与词条一起能够决定一种确切的词义，达到词义消歧的目的。对于机器翻译来说，词义消歧应该是一个非常重要的环节。但在我们所建立的 CLIR 系统中，在一定程度上，词义消歧对检索效率并未造成比较明显的影响。同时，为将更为充分的查询信息提供给检索系统，我们利用汉语《同义词词林》对经过翻译处理所获取的翻译知识进行扩展操作，即根据其中所描述的各种同义关系，罗列出与翻译知识相对应的各个同义词，从而达到查询扩展的目的。这样，不仅为检索系统提供更为丰富的查询信息，由此也大大提高检索效率，改善检索性能。

(2) 短语层翻译

该过程主要是利用英汉双语词典中的成语词典部分来实现，其中涉及到近距离与远距离短语识别的重要问题。这里，着重完成近距离短语的识别与翻译处理过程，所采用的方法为正向最大匹配法，描述如下：

- ✧ 从英汉双语词典中获取以当前查询词作为领头词的短语集合；
- ✧ 建立以当前查询词作为领头词，且其中所包含词汇数与所获取的短语集合中各成员所包含词汇数相同的各个短语；
- ✧ 比较所建立的各个短语及其所对应的短语集合中的各成员：
  - A. 如果有一对匹配成功，则加以短语标记，并将除已处理部分之外的第一个单词作为当前查询词，重复进行匹配过程；
  - B. 如果有多对匹配成功，则从中选取长度最大者加以短语标记，并将除已处理部分之外的第一个单词作为当前查询词，重复匹配过程；
  - C. 如果未匹配成功，则将当前查询词相邻的下一个单词作为当前查询词，重复匹配过程。

2.2 汉语 IR 系统

在经过翻译查询处理之后，利用汉语 IR 技术来获取相关文档列表。这里所采用的 IR 算法，是 MIT 方法[3]与概率方法的一种转换形式。其中，对于整个语料库的索引处理，我们利用最近几年所开发的汉语自然语言处理技术来完成[4]。而且，在处理过程中，同时采纳查询的标题域与描述域中单词的权重。

2.2.1 索引处理

在实现汉语 IR 的过程中，汉语文档的索引处理是最为重要的基础部分。由于汉语同英语显著不同，其词与词之间不存在空格。为成功地完成汉语文档的索引过程，必须首先对其进行分词处理，即将初始的汉语字符序列切分为单词或者 n-元组。为此，采用两种不同的索引方式对语料库中的文档进行处理，如下所示：

✧ 基于单词的索引方式

对于所给定的文档，首先必须根据文档所包含句子之间的标点符号（如句号、逗号等等），将其分割为句子序列。然后，利用句子层的分词处理器将每一个句子切分为单词序列。之后，利用文档后处理器对所获取的单词序列进行后续处理，主要是指单词序列中的错分以及漏分状况的检测，而最终生成分词结果。

✧ 基于 n-元组的索引方式

实现基于 n-元组的标注处理，其主要目的在于对基于 n-元组的 IR 方法与基于分词的 IR 方法进行比较，以寻求两者在效率上的差别。该索引方式主要是通过将文档简单地切分为 n-元组来完成，一般是指二元组，而不需要复杂的分词方法。

2.2.2 汉语 IR 算法

在所实现的 CLIR 系统中，所使用的汉语搜索引擎是基于统计的方法建立，并通过 TREC-5 中的汉语语料库进行调整。其中，该方法主要是利用由 MIT 语言系统研究组所提出的最大似然比作为文档的评价标准。

MIT 的研究者建议使用针对文档可能性的相对变化，将其表示作为条件概率与先验概率的似然比，从而对与查询 Q 相对应的文档进行评价并加以排序，如公式(1)所示。

$$S(D_i, Q) = \frac{p(D_i \mid Q)}{p(D_i)} \tag{1}$$

由此，语项权重可按照公式(2)进行定义，如下所示：

$$S_l(D_i, Q) = \sum_{t \in Q} q(t) \log\left(\frac{\alpha * p_{ml}(t \mid D_i) + (1-\alpha) * p_{gt}(t)}{p_{gt}(t)}\right) \tag{2}$$

其中，$q(t)$是查询中语项 t 的权重，一般是指其在查询中的频率。而

$$p_{ml}(t \mid D_i) = \frac{d_i(t)}{\sum_{t=1}^{k} d_i(t)} \tag{3}$$

其中，$d_i(t)$为语项 t 在文档 D_i 中出现的次数，k 为语料库中不同语项的数目，n 为文档集中文档的数目。

对于针对 $p(t)$的 Turing-Good 估计，通过 $p_{gt}(t) = p_r(t) = r^*/N$ 给出，其中

$$r^* = (r+1)\frac{N_{r+1}}{N_r} \tag{4}$$

r 为语项 t 在文档集中出现的次数，N_r为在文档集中出现 r 次的语项数目，N 为在文档集中所观察到的所有语项数目。

对于每篇文档，按照上述公式(1)-(4)，就可将其进行排序。

2.2.3 自动反馈与查询扩展

对于信息检索性能的改进，自动反馈已经证明为一种比较有效的方法。我们使用 MIT 方法的一种变体从引导搜索中选择语项。首先，如果 $\frac{S(D_i,Q)}{\max_{D_i} S(D_i,Q)} \le \gamma$，则选择文档 D_i，然后合并这些文档以创建一个联合文档 D'，如果文档 D'中的语项满足以下不等式：

$$\frac{p(Q'|D)}{p(Q')} \ge \frac{p(Q|D')}{p(Q)}$$

即，如果 $\frac{p_{ml}(t|D')}{p_{gl}(t)} \ge 1$，则选择语项，并将其加入至初始查询中，并赋予权重为 $-\log(\frac{p(t|D')}{p(t)})$，再次进行搜索。

然而，在所产生的结果中，许多文档与查询并无关系。通过对自动反馈过程做进一步研究，我们发现新近所加入语项的权重不应该与初始查询中单词的权重相等。因此，将初始查询中单词的权重设置为反馈单词的最大权重，从而对查准率与查全率进行优化。

3 实验测试

有关 CLIR 任务的评价任务是基于 TREC-9 所提供的英语查询集合以及三个汉语文档集来完成。前者包含 25 个英语主题，具有与其相对应的汉语翻译。而后者主要来自三个新闻集合，即香港商报、香港日报和大公报，总共包含 127,938 篇文档。

对于所建立的 CLIR 系统，我们以 TREC-5 中的汉语文档集作为训练集对其进行调整与优化，并以 TREC-9 汉语文档集作为测试集而获取四组运行结果：

- ✧ FDUT9XL1—以英语长查询（包含英语主题中标题域与描述域的内容）作为处理对象，并利用伪相关性反馈；
- ✧ FDUT9XL2—以英语长查询为处理对象，而未利用伪相关性反馈；
- ✧ FDUT9XL3—以英语中查询（仅包含描述域的内容）为处理对象，并利用伪相关性反馈；
- ✧ FDUT9XL4—以汉语长查询为处理对象，并利用伪相关性反馈，即单语 IR 运行结果。

在利用 TREC-5 汉语文档集进行训练的过程中，对于所获得的最好结果，其平均查准率(Mean Average Precision, MAP)可达到 0.3869。而在利用 TREC-9 汉语文档集进行测试的过程中，前三组运行采用自动查询翻译模式，其中利用基于分词的方法进行索引处理，而最后一组单语运行则采用基于 n-元组的切分方法完成索引处理。虽然最终的测试结果差于训练结果，但第二组运行“FDUT9XL2”仍可达到接近 0.30 的 MAP 值。有关四组运行结果的查全率/查准率曲线图以及相应的 MAP 值如图 1 所示。

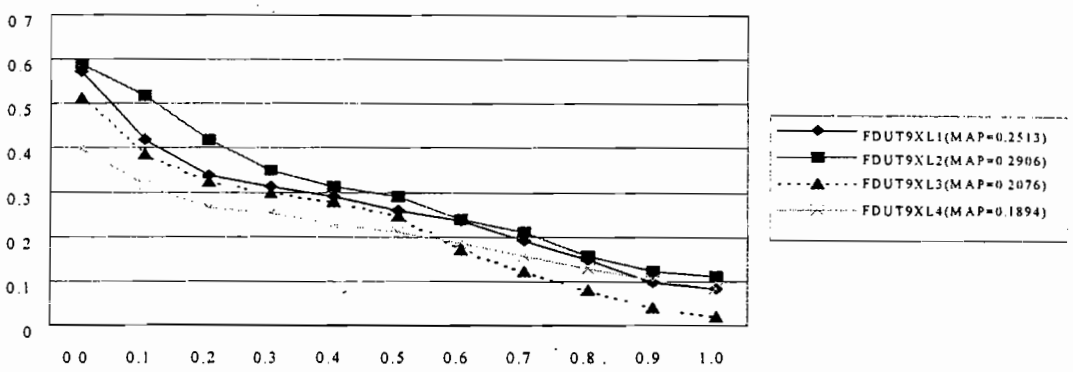

图 1、有关 CLIR 任务评价的查全率/查准率曲线图以及 MAP 值

从上图中可以看出，单语运行结果并不是很好。通过分析发现，主要归因于在 CLIR 处理过程中所使用的较为复杂的基于单词的索引处理方式，其性能优于基于 n-元组的索引处理方式，其中能够正确识别人名、地方名与组织名等等。而另一个因素，可能归因于在单语运行过程中所使用的翻译，即单语翻译所使用的单词与短语大都属于大陆风格，而与来自香港新闻报道的汉语文档集之间可能存在一些不匹配现象。

4 结论

本文以参加 TREC 中有关 CLIR 任务的评价作为研究背景，提出了一个面向英汉的跨语言信息检索系统的实现框架。其中，采取基于 MT 的查询翻译作为基本策略，并将注意力主要集中于两个方面：寻求实现英汉查询翻译的有效方法；以及寻求比较好的汉语 IR 策略。以所建立的 CLIR 系统为基础，我们获取了四组运行结果，其中包括三组 CLIR 运行以及一组单语运行，所获得的实验结果是令人鼓舞的。

参 考 文 献

[1] Mark W. Davis and Ted E. Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, editor, T*he Fourth Text Retrieval Conference (TREC-4)*. NIST, November 1995.

[2] Christian Fluhr. Multilingual Information Retrieval. In Ronald A Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Joe Zue, editors, *Survey of the State of the Art in Human Language Technology*, pages 291-305. Center for Spoken Language Understanding, Oregon Graduate Institute, 1995.

[3] Kenney Ng., "A maximum likelihood ratio information retrieval model", In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 1999

[4] Wu Li-de, et. al, Large Scale Chinese Text Processing (《大规模中文文本处理》), Fudan University Press, 1997.