

跨语言信息检索研究进展^{*}

张俊林 曲为民 杜 林 孙玉芳

(中科院软件所开放系统与中文信息处理中心 北京 100080)

摘 要 随着互联网资源的多语言性和用户所使用语言的日益多样性,跨语言信息检索成为越来越重要的研究领域。本文介绍了该领域为解决语言间的翻译障碍而研究的各种解决方法并且对汉英跨语言信息检索所面对的独特难点进行了分析,同时本文对国际上重要的跨语言信息检索系统评测会议进行了简单介绍。

关键词 跨语言信息检索,汉英跨语言检索,中文信息处理

State-of-the-Art : Cross-Language Information Retrieval

ZHANG Jun-Lin QU Wei-Min DU Lin SUN Yu-Fang

(Open System and Chinese Information Processing Center, Institute of Software, CAS, Beijing 100080)

Abstract Cross-Language information retrieval has become a very important research field because of the multilingual state of the Internet resource and Internet user. This paper introduces various technical approaches adopted in this field to resolve the translation problem and analyzes the special difficulties of Chinese-English Cross-Language information retrieval. Some important Cross-Language information retrieval system evaluation forums are also concisely introduced in this paper.

Keywords Cross-language information retrieval, Chinese-english CLIR, Chinese information processing

1 引言

随着信息化进程的全球化,互联网所提供的信息资源不再集中于英语等少数几种语言上。同时英语非母语的互联网用户比重也在快速增加,根据预测,到 2005 年,非英语互联网用户将增加到总用户数的 68%,其中中文用户增长速度最快,大约占总用户的 21%左右,而其它非英语语种也有不同程度的增长。对于大多数不精通外语的用户而言,熟练地使用外语查询所需要的相关的语种信息较为困难,而使用母语查询条件检索出相关的外语信息,再借助于辅助翻译工具浏览相关的信息则相对较为容易。因此自动将用户的母语查询条件翻译为相应的其它语种查询条件,再使用相关信息检索系统检索出所需的信息,是方便用户获取网上资源的有效途径。跨语言信息检索(CLIR: Cross-Language Information Retrieval)研究的正是这方面的内容,它是信息检索研究为了克服语言障碍而发展出来的一个分支。随着互联网的蓬勃发展,研究开发优秀的跨语言信息检索系统显得日益迫切。

跨语言信息检索系统的目标是在不同于用户查询条件语种的文档集中检索出与查询条件相关的文档。早期的跨语言信息检索可以追溯到 20 世纪 70 年代 Salton 建立的关于英德和英法德跨语言检索研究^[1],经过相关领域科研人员几十年的不懈探索,跨语言信息检索领域已经取得了很大的进展。本文对跨语言信息检索技术进行了系统的分析,以期对感兴趣的同行有一定的参考作用。我们首先介绍跨语言信息检索领域所采用的主要技术方法和路线并分析了汉英跨语言信息检索所面临的难点,然后简单介绍跨语言信息检索研究领域有

关的重要的语料测试会议,最后对跨语言信息检索研究进行总结和展望。

2 跨语言信息检索主要技术方法

跨语言信息检索的方法分类如图 1 所示。一般来说,解决查询条件与查询文档集之间的语言障碍有三种不同的技术路线^[2],既可以将查询条件翻译成与查询文档集相同的语言,也可以将查询文档集翻译成与查询条件相同的语言,还可以同时将两者映射到与具体语言无关的语义空间。由于查询文档集一般都很大,所以采取翻译查询文档集到查询条件语言来说代价太大,所以一般的研究集中在其它两种技术路线上。

2.1 受控词汇检索(Controlled Vocabulary)

受控词汇检索是指文档集通过手工根据预先选择好的词汇进行索引,用户也是从这相同的受控词汇中选择词汇构建查询条件,然后对文档进行检索。在该类检索系统中,多语言主题词表用于将被选择的每一种语种的词汇和与语种无关的概念识别器相联,文档的检索是通过概念识别器的匹配实现的^[3]。受控词汇跨语言检索系统中包含两个过程:将文档与查询条件都用受控词汇来表示。对文档的标识其实是对文档的翻译过程,而对查询条件用受控词汇来构建其实是对查询条件的翻译过程。它主要用于文档集的概念是可控的一些领域,例如图书馆中的文献检索。受控词汇检索的不足在于:

- 用受控词汇表中的检索词来标识每一篇文档通常是手工完成的,其使用范围受到很大的局限;
- 培训用户学会有效地使用受控词汇来构建查询条件是一件非常困难的工作。

^{*}基金项目:863 计划(No. 2001AA114040);国家自然科学基金(69983009)。张俊林 博士生,主要研究方向为信息检索,中文信息处理,Web 挖掘。曲为民 博士生,主要研究方向为机器翻译,中文信息处理。杜 林 博士,副研究员,主要研究方向为信息检索与中文信息处理。孙玉芳 研究员,博导,主要研究方向为系统软件与中文信息处理。

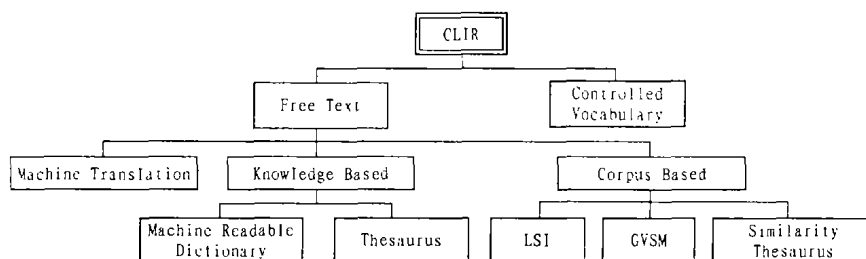


图1 跨语言信息检索方法分类图

2.2 自由文本检索(Free Text Retrieval)

对于文本检索,另一种不同于受控词汇检索的方法是用文档集中出现的语词来标识其中的文档,这种检索方法称作“自由文本检索”。跨语言信息自由文本检索的三个基本方法是:基于机器翻译的方法、基于知识的方法和基于语料的方法。

2.2.1 基于机器翻译的方法(Machine Translation Method)

实现跨语言检索系统最直接的方法是将机器翻译系统应用于检索过程中。一种方法是将用户查询条件翻译为与文档相同的语种^[4~6],另一种方法是将文档集中的文档翻译为与查询条件相同的语种,然后再用单语种的信息检索系统进行检索^[7]。

目前全自动机器翻译系统的主要缺点是翻译系统只能在受限的领域才具有较高的翻译质量。Fluhr 等人的研究表明,检索系统能较好地解决翻译中出现的语法错误,但无法解决翻译中出现的语义翻译错误,这主要是因为缺乏语词在不同领域的使用知识信息,因此将机器翻译应用到跨语言信息检索中会受到一定的限制^[8]。

对于较短的查询条件,由于没有上下文信息和领域信息,无法消除翻译中出现的歧义,其翻译质量会更差。而对于将翻译文档翻译成与查询条件相同的语种方法,即使可部分克服查询条件较短,缺少上下文的缺点,当翻译文档中出现一些错误可通过其他正确的翻译结果来弥补,但同单语种的信息检索系统相比,其检索性能要低很多。这主要是因为对于原文档中的多义词,在翻译过程中选择了最可能的语义进行翻译,而实际上也许并不准确,因此检索系统在翻译文档集上的检索性能要比在原文档集上的检索性能低。当原文档集所涉及的领域很广时,如互联网上的网页,全自动机器翻译的精度较低,更加影响了系统的检索性能。同时,目前全自动机器翻译系统的效率很难满足信息检索系统的要求。

2.2.2 基于知识的方法(Knowledge Based Method)

基于知识的方法包括基于机读双语词典的方法和基于主题词表的方法。

基于机读双语词典的方法(Machine Readable Dictionary Method)。双语词典原本是为了方便人类自身学习外语而构建的,现在随着电子版本的流行,它已经是用于跨语言信息检索中最广泛的知识结构,基于双语词典的查询条件的翻译方法也是跨语言信息检索领域中最为关注的方法^[9~11]。

基于双语词典的查询条件翻译方法的基本思想是将查询条件的检索词用恰当的目标语种的一个或者一组检索词来代替。该方法主要有三个方面的因素需要考虑。一是许多检索词在基于双语词典的翻译过程中,它们的译项并不是唯一的,而且这些译项中所表达的意思也可能根本不相似。在单语言信息检索过程中也会遇到同义词和一词多义的问题,但翻译过程中的歧义更加剧了问题的严重性。第二个需要考虑的因素

是用于翻译的双语词典中可能不包含源语种检索词,也就不包含它的译项,而这些检索词可能是用户的查询条件关键检索词。这可能是由于用户在查询条件中包含了一些专有名词或者是缩略语造成的。第三个影响检索效果的因素在于一般的双语词典中只有很有限甚至没有关于短语的翻译项。对于“非组合短语”(Non-Compositional Phrase)来说这个问题尤其突出。所谓“非组合短语”,是指下面两种情况:一种是有些短语如果拆开逐词翻译,不可能得到正确的译文,其原因是双语词典中每个词的候选译词无论怎样组合都不可能得到正确的短语翻译。更多的情况下一个概念是以短语的形式表示的,如果把组成该短语的词汇拆开后,每一个词的语义叠加起来不等于该短语的语义。对于非组合短语,按照基于双语词典的方法进行逐词翻译,不可能或者很难得到准确的短语翻译,进而不可能进行正确的检索。

目前很多研究集中在如何解决使用双语词典所面临的这三个主要障碍上面。在解决翻译项的歧义方面,通常使用的方法是:1. 基于语法分析的方法;2. 基于统计的方法;3. 基于查询扩展的方法。Davis 在对 25 个西班牙语的查询条件利用 Collins 双语词典翻译成英语时提出了三种消除翻译歧义的方法^[12]。第一种是利用词性信息来约束翻译过程。第二种是利用单语句与双语句平行的语料选取恰当的译项。第三种是前两种方法的结合,并得到最好的检索性能,大约为单语言检索性能的 73.5%。Ballesteros 和 Croft 在对 25 个西班牙语的查询条件翻译前进行查询扩展与翻译后进行查询扩展的策略来消除翻译歧义^[13]。他们分别对翻译前进行查询扩展,翻译后查询扩展和前两者的结合方法进行了实验,试验表明第三种方法的检索性能最好,大约为单语言检索性能的 68%。在解决词典词汇覆盖不足问题上面,Hiemstra 提出了基于 3-gram 和编辑距离来将查询语词与索引项进行模糊匹配的方法^[14],这个方法对于解决词典词汇覆盖不足以及减少用户输入查询条件时拼写错误都有很好的效果。在短语识别与翻译方面,研究人员也想出了很多方法比如翻译图^[15]、自动机^[16]等方法来对查询语句中识别出来的名词短语进行正确识别和翻译。

基于主题词表的方法(Thesaurus Based Method)。主题词表是管理词语复杂性的一种结构并提供语词间概念的联系,它能从语义的角度说明词语的含义,是概念与术语的知识库^[17]。

主题词表还能够对领域知识进行管理。每个多语言主题词表都具备的一个关键特征就是跨语言同义词规范,用于跨语言信息检索的主题词表由于引入了跨语言同义词规范,使得不同语种的主题词之间能够相互对照。在复杂一些的主题词表中还包含概念结构信息以及词汇或者概念间的反义、关联词等相互关系。

Salton 等人在 1969 年首次进行了跨语言的检索试验,他

们在 SMART 检索系统中,将英语检索使用的概念列表翻译为德语,构造相应的德语概念列表,同时手工将 48 个英语查询条件翻译为德语,使用英语查询条件检索 468 篇德语摘要,与德语查询相比,系统的查准率从 0.35 下降为 0.34。类似地,使用德语查询 1095 篇英文摘要,系统的查准率从 0.33 下降到 0.31。根据这样的试验结果,他得出结论“跨语言检索性能与单语言检索性能相近”,但强调要不断完善主题词表。1973 年,Salton 构造了较为全面的英法跨语言概念列表,使用 52 篇摘要作为平行语料,并翻译了 16 个查询条件,使用英语查询条件检索法文摘要时比用法语查询条件检索查准率提高了 5%,而使用法语查询条件检索英文摘要时比用英语查询条件检索查准率降低了 12%。EMIR(European Multilingual Information Retrieval)项目利用快速浅层分析来得到语法信息并将它加入主题词表中。通过对英语检索词进行词性标识,然后再选择句法上与法语检索词最恰当的英语检索词作为主题词。EMIR 主题词表是英法双语词表,语词间的关系用符合词来表示。在英法跨语言试验中,基于 EMIR 主题词表的系统查准率为 0.27,而基于机器翻译软件 SYSTRAN 的系统查准率为 0.21,提高了 29%^[17,18]。

由于主题词表以人类思维的方式来表示概念或术语间的联系,因此它能使用户构建出更好的查询条件。另外由于领域中的许多知识能够加入到主题词表中,对于熟悉主题词关系的用户来说基于主题词表的信息检索是一个强有力的工具。然而主题词表对用户使用的语汇和检索系统所使用的领域具有限制,而且主题词表的构建和维护需要大量的人力和物力,要有效使用复杂的主题词表中概念的关系需要对用户进行大量的培训才能掌握。

2.2.3 基于语料的方法(Corpus Based Method) 由于在现实生活中不可能构建完备的双语词典或是手工构建复杂的主题词表,因此基于语料的方法从分析现有大规模的语料入手,从中抽取所需信息自动构建与应用有关的翻译技术^[19]。这些语料分为平行语料(语料中包含文档及其相应的翻译文档,按照文档翻译的方式又分为文档对齐、语句对齐和语词对齐三种方式)和可比语料(语料中包含不同语种的涉及相似主题的文档,不同语种的文档之间不存在一一对应的关系)。

基于平行语料的方法(Parallel Corpus)。对于文档对齐的平行语料,Landauer 和 Littman 提出了跨语言潜在索引(CL-LSI:Cross-Language Latent Semantic Indexing)的信息检索技术^[20~23]。它的基本思想是首先通过将有代表性的文档与其对应的翻译文档联系起来形成训练文档集,然后利用奇异值分解技术(SVD:Singular Value Decomposition)对双语检索词-文档关联矩阵进行奇异值分解,获得双语文档集的特征信息以及检索词用法上的映射关系,即构造出不同语种的潜在语义空间,最后根据平行文档中语词的用法特征可检索出另一种语种的相关信息。

伪相关反馈方法(Pseudo-relevance Feedback)也是基于平行双语语料的一种方法。首先源语查询条件在平行语料中的源语文档集合中进行单语检索,从检索返回结果中选择一定数量的前列结果并找到与这些文档一一对应的目标语文档集合,从这些目标语文档集合生成目标语查询条件^[24],然后在待查的目标语文档集中进行检索,所得到的结果就是跨语言检索的最终结果。

另外一种平行语料方法是广义向量空间模型(General-

ized Vector Space Model,GVSM)^[25],它的基本思想是根据双语训练文档集分别建立源语与目标语的检索词-文档关联矩阵,在计算查询条件和文档的相似度时,考虑将经典的向量空间模型与两个关联矩阵相结合在源语与目标语之间实现映射关系。

基于可比语料的方法。构建平行语料较为昂贵,因此也较难得到,而可比语料则要广泛得多,它对于文档的对齐只限于主题相似即可,但由于现在自动或者半自动对齐技术与应用密切相关,所以该方法仍受到一定的限制。

使用可比语料最著名的方法是相似主题词表方法,Sheridan 等人为了克服平行语料的限制,根据新闻事件的内容和时间构造了德语和意大利语可比语料库。他们使用可比语料中抽取出来的跨语言相似主题词对查询条件进行扩展,将德语查询条件转换为意大利语对应的查询条件进行检索。同基于跨语言的潜在语义索引相类似,跨语言相似主题词是根据检索词在可比语料中的分布信息确定的。实验结果显示跨语言信息检索的性能为单语言信息检索性能的一半左右^[26],同其它基于语料的跨语言信息检索方法类似,基于可比语料的跨语言信息检索方法也与语料领域相关。

3 汉英跨语言信息检索难点分析

对于汉英跨语言信息检索来说,除了面临一般跨语言信息检索所具有的难点外,还有下列一些与其语言有关的不利因素。这些不利因素主要还是由于汉语的一些特性所导致的。

1. 中文文本之间没有分隔符,在基于检索词的检索系统中必须要对其进行中文语词的切分,中文语词的切分算法的有效性直接影响到查询条件的翻译效果。

2. 至今仍没有较好的方法识别汉语中的新词、外来词、专有名词、人名、地名和缩略语等语词,自动构建汉英双语词典较为困难,这势必影响查询条件的翻译效果。这些语词也是第一点中基于中文词语的切分方法所面临的困难。

3. 汉语没有词级变化,并且语词的词性兼类较多,这给基于汉英双语词典的查询条件翻译方法带来许多困难。

4. 汉语句法、语义分析复杂,至今还没有较好的研究成果运用于中文信息检索领域。

4 重要的跨语言检索系统评测会议

传统对于信息检索系统的评测都在标准化的实验室环境中进行,以比较检索系统或者检索技术的检索性能,然而早期的测试文档集规模通常较小,与真实的检索环境之间有很大的差距,因此基于这样的测试集的检索系统,在实际应用中无法达到较好的性能。对于象跨语言信息检索这样很大程度上依赖实验的学科,相关的检索系统评测会议对于该学科的发展起到了很大的推动作用。跨语言信息检索方面比较有影响力的系统评测会议包括 TREC、CLEF 和 NTCIR 评测会议。

TREC(Text Retrieval Conference Series)。在 1992 年,美国国防部高级计划研究计划局(DRAPA)与美国国家标准技术局(NIST)共同举办了文本检索会议(TREC)。通过大型测试文档集的建立,测试项目、测试程序、评估准则的制定,为不同的检索系统和检索技术提供了一个标准的评比环境,并举办论坛为参与者讨论和分享结果。它首创了前所未有的大型测试文档集,使得测试环境得以更加接近真实情况,对检索技术的发展与系统性能的提升具有很大的贡献。到目前为止已经举办了 11 届,从第 6 届 TREC 会议开始,CLIR 成为正式

的测试项目之一。TREC 会议主要集中在西方语言之间的跨语言检索,后来也增加了中文、阿拉伯文与英文之间的跨语言检索评测项目。

CLEF(Cross Language Evaluation Forum)。CLEF 是欧洲委员会资助的数字图书馆研究中的一部分研究内容。它是与欧洲语言跨语言信息检索有关的评测会议。从 2000 年开始到目前为止已经举办了三次会议。其测试项目包括欧洲语言的单语检索,跨语言与多语言检索,受限领域检索以及交互检索。涉及的欧洲语种有十多种。

NTCIR(NACSIS Test Collection for Information Retrieval)。NTCIR 是由日本国立信息研究所(NII)主办的信息检索测试集评测会议。它主要构建亚洲语言有关的测试文档集以及对相关的检索系统做出评测。到目前为止已经成功地举办了三次 NTCIR 会议。测试项目包括了单语言信息检索,跨语言信息检索,问答系统,文本摘要以及 Web 检索。其中跨语言信息检索涉及到的亚洲语言包括中文、英文、日文和韩文。

结论与展望 本文重点叙述了跨语言信息检索领域所采用的主要技术方法,分析了该领域所面临的一些难点并对汉英跨语言信息检索所面临的独有难点进行了说明,同时介绍了该领域比较有影响的国际评测会议。近些年来跨语言信息检索领域无论在研究方法还是在测试集的构建方面都取得了长足的进步。实验结果表明^[27~30],多种方法结合的效果要优于单种方法。所以该领域一个很明显的趋势就是越来越多的研究人员开始考虑结合本文提到的上述方法中的几种来进一步改进查询翻译的精度。我们认为如何更好地结合多种现有手段的综合方法研究将会成为跨语言信息检索的研究热点。

尽管在该领域的研究有很大的进展,但是在真实语料环境下的可用跨语言检索系统及其相关工具还是非常缺乏,现有技术离真正的实用化还有一段距离,这对该领域的研究者来说无疑仍是一种比较大的挑战。

参考文献

- Oard, Douglas W. Alternative approaches for cross-language text retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval, 1996. <http://www.glue.umd.edu/oard/research.html>
- McCarley J S. Should we translate the documents or the queries in cross-language information retrieval?. In: 37th Annual Meeting of the Association for Computational Linguistics College Park, MD, 1999
- Soergel D. Multilingual thesauri in cross-language text and speech retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997
- Gey F C, Jiang H, Chen A. Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire 2 at TREC-7. In: Proc. of the 7th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 1998
- Gareth J, et al. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In: Proc. of Cross Language Evaluation Forum, 2001
- Kowk K L. English-Chinese cross-language retrieval based on a translation package. In: Proc. of the 22th Intl. Conf. on Research and Development in Information Retrieval, 1999
- Oard D W, Hackett P. Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: Proc. of the 6th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 1997
- Fluhr C. Multilingual information retrieval. Survey of the State of the Art in Human Language Technology, Center for Spoken Language Understanding, Oregon Graduate Institute, 1995. 391~405
- Chen A, Jiang H, Gey F. English-Chinese Cross-Language IR using Bilingual Dictionaries. In: Proc. of the 9th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 2000
- Gao Jianfeng, et al. TREC-9 CLIR Experiments at MSRCN. In: Proc. of the 9th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 2000
- Hull D A, Grefenstette G. A dictionary-based approach to multilingual information retrieval. In: Proc. of the 19th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'96), 1996
- Davis M. New experiments in cross-language text retrieval at NMSU's computing research lab. In: Proc. of the 5th Text Retrieval Conf. (TREC-5), 1997. 391~407
- Ballesteros L, Croft W B. Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proc. of the 20th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1997. 84~91
- Hiemstra D, Kraajj W. Twenty-One at TREC-7: Ad-hoc and Cross-language track. In: Proc. of the 7th Text Retrieval Conf. (TREC-7) 1999
- Kraajj W, Hiemstra D. Cross Language Retrieval with the Twenty-One System. In: Proc. of the 7th Text Retrieval Conf. (TREC-6) 1998
- Chen A, et al. Berkeley at NTCIR-2: Chinese, Japanese and English IR experiments. In: Proc. of the second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization, 2001. 137~146
- Radwan K, Fousier F, Fluhr C. Multilingual access to textual database. In: Lichnerowicz A, ed. In: Proc. of a Conf. on Intelligent Text and Image Handling (RIA0 91), 1991. 475~489
- Radwan K, Fluhr C. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In: Fourth Annual Symposium on Document Analysis and Information Retrieval, 1995. 121~136
- Nie J Y, et al. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. ACM-SIGIR Conf. Berkeley, CA, 1999. 74~81
- Landauer T K, Littman M L. Fully automatic cross-language document retrieval using latent semantic indexing. In: Proc. of the 6th Annual Conf. of the UW Center for the New Oxford English Dictionary and Text Research, 1990. 31~38
- Landauer T K, Littman M L. A statistical method for language-independent representation of the topical content of text segments. In: Proc. of the 11th Intl. Conf. : Expert Systems and Their Application, 8, 1991. 77~85
- Deerwester, Dumais S T, Furnas G W, Landauer T K, Harshman R A. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391~407
- Dumais S T, Landauer T K, Littman M L. Automatic cross-linguistic information retrieval using latent semantic indexing. SIGIR96 Workshop On Cross-Linguistic Information Retrieval, 1996
- Rogati M, Yang Y. Cross-lingual pseudo-relevance feedback using a comparable corpus. In: Proc. of Cross Language Evaluation Forum 2001
- Wong S K M, Ziarko W, Wong P C N. Generalized vector space model in information retrieval. In: Proc. of the Eighth Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, Association for Computing Machinery, 1985. 18~25
- Sheridan P, Wechsler M, Schauble P. Cross-language speech retrieval. In: Proc. of the 20th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1997. 99~109
- Braschler M, et al. The Eurospider Retrieval System and the TREC-8 Cross-Language Track. In: Proc. of the 8th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 1999
- Sadat F, et al. Query Expansion Techniques for the CLEF Bilingual Track. In: Proc. of Cross Language Evaluation Forum 2001
- Kwok K L, Grunfeld L, Dinstl N, Chan M. TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS. In: Proc. of the 9th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 2000. 419~423
- Gey F C, Oard D W. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries. In: Proc. of the 10th Text Retrieval Evaluation Conf. National Institute of Standards and Technology, 2001. 16~20