Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

多语主题词表及其应用研究*

□ 徐红姣 张均胜 王惠临 / 中国科学技术信息研究所 北京 100038

摘要：对多语言信息进行语义层面的精确描述，为用户提供准确的跨语言信息资源，是当前多语言信息服务中必须面临和解决的实际问题。多语言主题词表正是解决这一问题的有效工具资源之一。文章首先介绍了国外三个常用多语言主题词表，然后对多语言主题词表在多语言信息自动标引和多语言信息检索两个领域中的应用情况进行了分析，说明多语言主题词表在多语言信息服务领域的潜在应用价值。

关键词：多语言主题词表，多语言自动标引，多语言信息检索
DOI：10.3772/j.issn.1673—2286.2013.12.007

1 引言

互联网技术的发展使得人们能够摆脱地域限制进行交流，但随着网络上不同语言信息资源的日益增加，人们越来越多地面临着如何利用多语言信息的问题，迫切需要能够适用于多语言信息处理的相关工具、方法和技术。为用户提供具有丰富语义信息的多语种信息资源成为当前多语言信息服务必须解决的实际问题。多语言主题词表正是解决这一问题的有效资源之一。

主题词表由词及词间相互关系组成，能反映各学科领域主题词间的语义关系，是结构化的概念集合，在信息资源的描述、组织和检索中发挥重要的作用。多语言主题词表在普通主题词表的术语及关系中，加入了不同语种的映射。这使得它成为多语言信息组织和检索的重要工具，而且作为多语种的语义词典，其在语义网、跨语言知识组织与管理等方面均有广阔的应用前景。

本文对国外多语主题词表及其应用现状进行了初步的研究，首先介绍了三个常见的多语主题词表，然后对多语主题词表在多语言自动标引和跨语言信息检索两方面的应用现状进行了探讨，说明多语主题词表在多语言信息服务中的应用价值。

2 国外常用多语主题词表

国外非常重视多语主题词表的研究与应用，已有大量实用的多语主题词表，下面我们将简单介绍三个常见的多语主题词表。

2.1 EuroVoc

EuroVoc[1]，即欧盟主题词表，是欧洲议会、欧盟委员会和欧盟出版局于1982年开始开发的多语言、多领域主题词表，其目标是为信息管理和传播服务提供一个一致的标引工具，实现对文档资源的有效管理，帮助用户进行基于受控词的文档检索。最新版本的EuroVoc于2012年年底发布，涉及25个语言版本（包括22个欧盟语种），包含英文叙词6883个，非叙词8348个，涵盖政治、经济、金融、科学、运输等21个领域。从2000年底起，普通用户可以通过EuroVoc的官网来浏览和查询其最新版本，也可以按领域或者字母顺序下载PDF版本的EuroVoc，或者下载不同语种的主题词对应的excel文件。EuroVoc主题词表在欧洲的应用范围比较广，包括EUR-Lex（http://eur-lex.europa.eu/）、EU Bookshop （http://bookshop.europa.eu/）、欧洲议会、欧盟多个组织和机构、欧洲多个国家和地区的议会和政府部门等。

2.2 AGROVOC

AGROVOC[2]是由联合国粮农组织和欧盟委员会

* 本文系中国科学技术信息研究所2012年预研项目“基于主题词表的多语言科技信息组织与检索方法研究”（编号：YY201223）及重点工作项目“多语言科技信息语义关联网络构建及其应用”（编号：ZD2012-3-3）的研究成果之一。

于上世纪80年代开发的农业多语言主题词表。但是它的领域不仅仅限于农业，还覆盖了食品、农业、渔业、林业、环境等诸多领域。目前AGROVOC有22个语种版本，包含30000多个概念。作为一个重要的多语种叙词表，AGROVOC广泛应用于农业领域信息资源的组织和检索，世界三大农业数据库之一的AGRIS数据库就是利用AGROVOC进行信息组织的。为了方便用户使用，FAO官网免费提供最新版本的AGROVOC浏览、搜索和下载服务，下载的格式包括MySQL、Microsoft Access、XML、OWL等。同时为了鼓励农业信息管理系统开发人员将AGROVOC应用到他们的系统中，FAO还提供AGROVOC的在线服务，免去了用户必需经常进行下载的不便，也保证使用最新版的词表。

2.3 MeSH

MeSH[3]（Medical Subject Headings，医学主题词表）是美国国立医学图书馆（NLM）编制的、生物医学领域广泛使用的最为权威的大型综合性主题词表，其目前已被翻译成包括中文在内的20多个语种。MeSH于1960年首次出版，此后NLM每周都会进行更新，每年都发布新版本的MeSH。自2007年起，NLM停止了印刷版MeSH的出版。用户可以从NLM的网站上下载英文版的MeSH的主题词、副主题词、补充概念及树结构等，下载形式有XML、ASCII、MARC等。

MeSH Browser[4]是网络版MeSH检索系统，能够帮助用户查找主题词、副主题词和补充概念等，查看完整MeSH记录，显示概念间的相互关系。为了方便对MeSH的翻译，NLM开发了MeSH翻译维护系统MTMS，实现对翻译结果的持续更新，同时追踪MeSH每年的更新，以便及时对其他语种版本的MeSH作出修改[5]。

表1对上文所述的三个多语主题词表进行了概括。

除了上述列举的三个主题词表外，国外还有大量的多语言主题词表，例如联合国教科文组织开发的英、法、俄、西班牙语UNESCO主题词表，欧盟教育领域英、法、德语主题词表EUDISED，欧洲文化遗产网（European Heritage Network）开发的文化领域多语叙词表HEREIN，欧洲环境总署开发的近13个语种的环境领域主题词表GEMET等。美国国会图书馆主题词表LCSH也被法、德、希腊、匈牙利、波兰等多个国家以翻译、与本国现有主题词表映射等多种方式形成多语主题词表，应用到不同信息服务系统中。相比起来，国内的多语言主题词表研究较为滞后，为了能够更好地利用多语言信息资源，开展对多语言主题词表的研究是非常必要的。

3 基于多语主题词表的多语言自动赋词标引

多语主题词表能够克服语种的限制，将不同语种的文档纳入到统一的知识系统，揭示它们的相互关系，

表1 三个常用的多语主题词表

语 种	主题领域	规 模	语 种	版权机构	应 用
EuroVoc	综合	英文叙词6883个，非叙词8348个（4.4版）	25	欧盟	EUR-Lex、EU Bookshop、欧洲议会、多个国会图书馆（意大利、西班牙、捷克、斯洛伐克）等
AGORVOC	农业	32188个概念 626211个术语（2013版）	22	联合国粮农组织	AGRIS、National Agricultural Library（美国）、Canadian Agriculture Library等
MeSH	医学综合	26853个概念，对应213815个术语；83个副主题词；209420个补充概念，对应510687个术语 （截至2012年9月4日）	21	美国国立医学图书馆	MEDLINE、PubMed、DDRT（瑞典）、OMNI（英国）等

实现多语言文档集合的有序化。传统的基于主题词表的信息资源标引主要是依靠手工进行，费时费力。为了克服人工标引效率偏低且不能满足一致性要求的缺陷，学者们开始研究自动标引技术。基于多语主题词表的自动赋词标引方法大致可以分为三类：基于语言分析的方法、基于统计的方法及混合方法。下面将对这三种方法进行介绍。

3.1 基于语言分析的方法

多语言自动赋词标引的对象是不同语种的自然语言文档，因此人们便尝试从语言学的角度进行多语文档的自动标引。基于语言分析的多语言自动标引方法主要利用词形还原[6]、复合词分解[7]、去除停用词、短语/组块识别[8]等方法分别对待标引的文档和主题词表中的叙词进行处理，将处理后的文档中的词汇和主题词表中叙词进行机械匹配，为文档赋予标引词。

主题词表中的词汇为受控词汇，通常会与文档中的自然语言词汇有很大的形态差异，利用各种语言学方法可以很好地弥补两者间的差异。但是，此方法没有考虑文档和主题词表中叙词的语义关联性，生成的标引词通常都在文档中出现过，仅仅是词形上同主题词表中的术语有所不同。在大多数研究中，此种方法都作为对文档的预处理手段，结合其他统计或机器学习方法来提高自动标引的准确率。

3.2 基于统计的方法

作为信息组织的重要工具，主题词表广泛应用于各种文献资源的标引，积累了大量的人工标引的多语种文献资源。基于统计的多语自动赋词标引方法分析现有的已经标引好的文档资源，从中获取标引模型并将其应用于新的文档的自动标引。标引模型的获取可以通过简单的统计算法。例如，文献[9]中利用人工标引的文档，通过log-like计算获取每个叙词的相关词，然后通过统计TF、DF、文档的标引词个数、叙词在训练语料中出现的频率等参数，计算叙词和相关词的关联度，从而得到标引模型。在标引阶段，对于每个待标引文档，依据相同的方法获取候选标引词及其权重，然后计算文档的候选标引词向量和每个叙词的相关词向量的相似度，相似度较高的叙词即可作为最终标引结果。此方法受训练语料规模的影响，语料规模较小时，某些叙词始终没有作为标引词出现在文档集中，无法获取其特征向量，从而影响自动标引的结果。

随着人工智能技术的发展，机器学习算法也越来越多地应用到多语自动标引中，例如文献[10]和文献[11]中分别利用贝叶斯网络和SVM算法建立自动标引模型，取得了较好的自动标引结果。机器学习算法的应用，在一定程度上提高了自动标引的准确率，但还是存在着数据稀疏和关键词漏标等问题。总体来说，基于统计的多语自动标引方法不受语种的限制，不必针对不同语种的文档确定不同标引方法，非常适合多语种文档的自动赋词标引。

3.3 混合方法

混合方法也即将语言学知识和统计特征结合起来进行多语言文档自动赋词标引的方法。尽管基于统计的多语自动标引方法能够不受语种限制，但是正如上文所述，多语主题词表中的受控词汇和文档中的自然语言词汇存在着很大差异，语言分析是多语言自动标引不可缺少的步骤。Bruno Pouliquen等人的研究结果证明了对语言分析的重要性：通过使用词形还原、多词短语标注和去除停用词三种方法，英语的自动标引准确率由45.6%上升到50%，西班牙语的准确率则由40.3%上升到了46.2%[9]。因此目前大部分的研究成果都采用先用语言分析方法处理文本获取候选标引词，再利用统计模型确定标引词的多语言自动标引模式。

国外已有很多实用的基于多语主题词表的多语言自动赋词标引工具，表2中列举了三个典型的基于多语主题词表的多语言自动标引工具并对它们进行了简单的介绍。

3.4 面临的问题

尽管多语言自动赋词标引技术已经有了很大的发展，但是还存在着很多的问题，主要体现在：

（1）没有充分利用主题词表中丰富的语义关系。相比于普通的词表，主题词表的最大特点就是蕴含了大量的概念间的等同、等级和相关关系。常见的主题词表词间关系的利用方式是从文档中选取候选标引词时将候选标引词扩展到其所有相关词[11,15]，文献[16]中利用贝叶斯网络对EuroVoc主题词表中的概念及概念间关系进行建模，利用概率推理算法为文档进行赋词标引。多

表2 基于多语主题词表的多语言自动赋词标引工具

名 称	开发机构	针对的多语主题词表	主要技术特点
JEX[12]	JRC（欧盟委员会联合研究中心）	EuroVoc	利用统计机器学习算法从已标引好的文档集中学习多标签分类规则
Agrotagger[13]	IITK（印度Kanpur技术研究所）、FAO、ICRISAT、MIMOS	AGROVOC	识别文档中所有的Agrotags词（Agrotags为AgroVoc的子集）并利用统计算法计算其作为标引词的概率
KEA[14]	The University of Waikato	任意多语主题词表	利用朴素贝叶斯等算法从已标引的文档集中学习关键词抽取策略并应用于新的文档标引

语自动标引的目标不仅是要揭示文档的主题内容，更为重要的是确定不同语种间文档的相互关系，因此需要充分地利用主题词表提供的丰富语义关系来提高自动赋词标引的准确率。

（2）自动赋词标引的准确率不高：虽然自动标引技术多种多样，但由于技术的限制，小规模试验的效果较好，大规模应用的标引质量还是不高。因此目前自动赋词标引技术一般用来辅助进行人工标引，帮助提高人工标引的速度，克服人工标引一致性较差、随意性较大等缺点。正如Lancaster等人所说，自动标引技术距离完全实际应用仍有很长的距离，只有机器具有足够智能，才能完全替代人类完成这一工作[17]。

（3）自动赋词标引的评价问题：传统的自动标引评价是对照人工标引结果判别或者由专家打分，这种方法主观性大，成本也比较高。人工标引和自动标引的特点不同，自动标引中专注于文档中的词汇描述，一般给出的标引词都比较具体，而人工标引中，考虑到用户的使用习惯，通常都会使用比较概括的词作为标引词[18]。鉴于人工标引与自动标引的不同特点，构建一个适用于自动赋词标引的评价模型是一项很有价值的研究工作。

4 多语主题词表在多语言信息检索中的应用

多语言信息检索是利用一种语言的查询式检索多种语言文档集合的技术，主要需要解决的问题是如何将不同语种的查询式和文档进行匹配。多语主题词表刻画了不同语言中对应的领域知识，从而更好地解决从源语言到目标语言之间转换过程中出现的语义损失和曲解等问题，有效地理解用户的查询意图，获得预期的检索结果。在早期的多语言信息检索研究中，基于多语主题词表的方法占据了主导地位，而目前大多数实用的多语言信息检索系统也都或多或少地应用了多语主题词表，表3列举了三个基于多语主题词表的多语言信息检索项目。

多语主题词表在多语信息检索中应用机制主要有三种：

（1）在查询翻译过程中，利用主题词表概念、概念间关系等信息进行的语义层面的翻译，克服由于缺乏语境造成的查询翻译不准确问题，实现查询翻译消歧；

（2）在查询式翻译前或者后，利用多语主题词表中的上下位、相关关系等对用户提交的查询式或者查询式的翻译结果进行扩展；

（3）对文档建立基于概念的索引，通过语义分析得到揭示文本内容的标引词，过滤文本存在的语义歧义，提高检索的准确率。

第三种应用机制就是利用多语主题词表对文档进行标引，上文已有详细介绍，下面我们将对多语主题词表在查询翻译和查询扩展中的相关研究进行介绍。

4.1 多语主题词表与查询翻译

主题词表在查询翻译中的应用有很长的历史。1970年Salton进行的第一次跨语言信息检索实验使用的翻译资源就是多语言主题词表[19]。多语主题词表提供了不同语种的主题词之间的相互对照关系，还包含了

表3 三个基于多语主题词表的多语信息检索项目

项目名称	研发机构	支持语种	使用的多语主题词表	主题词表的应用机制
MuchMore	CMU、DFKI、Stanford、Eurospider、XEROX等	德语、英语	UMLS	利用UMLS等语义资源对查询式和文档进行语义标注，以语义标注结果为媒介，对不同语种的查询式和文档进行匹配
TRANSLIB	KNOWLEDGE S.A.、University of Patras等	英语、希腊语、西班牙语	EUROVOC	利用EUROVOC对用户输入查询式进行控制，实现多语言受控检索
LAURIN	University of Innsbruck、University of Roma等	英、法、德、意大利、挪威、西班牙、瑞典语	LAURIN Thesaurus	支持多语主题词表的构建、在线维护和更新，基于LAURIN Thesaurus的多语标引、受控检索

主题词结构信息以及与之相关的概念信息，可以应用于对查询式进行翻译和翻译消歧。

将多语主题词表作为查询翻译资源使用的过程中需要解决的一个重要的问题就是如何将主题词表中的受控词汇和用户查询式的自然语言词汇进行匹配。有些多语主题词表本身就蕴含了大量的概念和词汇间的映射关系，可直接应用到查询翻译过程中。例如，UMLS超级叙词表采用概念-术语-词串三级结构模式，将一个概念的不同术语连同术语的多个变异词串有序地组织到一起，David Eichmann等人[20]就利用UMLS的这种结构模式对查询式进行翻译。对于没有提供词汇和概念间映射关系的主题词表，可以采用将查询式中的词汇映射到主题词表的概念的方式进行查询翻译。Julio Gonzalo等人[21]采用词形标注、短语识别、语义距离计算等方法将查询式映射到EuroWordNet的中间语索引（InterLingual Index）中，实现概念层面的查询式翻译。

查询翻译过程中面临的最大问题就是歧义问题。自然语言中一词多义和一义多词的现象非常普遍，而用户查询式的长度通常都很短，要准确判断词的含义就很困难。主题词表中包含的丰富概念和概念间关系信息为消除歧义提供了很好的语境信息。Yarowsky[22]提出了一种利用Roget词表中每个义类中所有词的上下文信息确定一个多义词义的方法。用这种方法对英语中12个多义词进行义项标注，平均准确率达92%。Ahmad M. Hasnah等人[23]利用多语主题词表中的同义词和相关词，对基于双语词典的查询式的翻译结果进行消歧。作为一种重要的语义资源，主题词表在词义消歧方面有着很重要的作用。

多语主题词表应用到查询翻译中，最主要的问题就是其覆盖度问题。主题词表一般都是面向某个特殊的领域，提供的翻译信息也仅限于此领域。但是用户输入的查询式中的词汇不可能全部都是该领域中的词，仅仅利用领域主题词表无法取得很好的翻译。可以采用将特定领域主题词表和通用领域的主题词表相结合的方法[24]，或利用其他领域平行或可比语料库弥补现有主题词表覆盖度不够全面的问题[21,25]。

4.2 多语主题词表与跨语言查询扩展

跨语言查询扩展按照查询扩展发生在跨语言信息检索过程中的先后顺序不同可以分为翻译前查询扩展、翻译后查询扩展以及两者的结合。跨语言查询扩展和单语言查询扩展在本质上没有区别，只是将单语言查询扩展的方法运用到跨语言信息检索过程的不同步骤，本质还是单语言查询扩展原理。单语言查询扩展方法分为两种：基于全局分析和基于局部分析方法，其中基于全局分析的方法通常就利用主题词表、同义词词典等工具进行查询扩展，因此多语主题词表也经常用于跨语言查询扩展中。

利用多语主题词表进行跨语言查询扩展，最简单的方式是利用主题词表的层次结构，直接将主题词表中与查询式相关的主题词的同义词、上位词、下位词或相关词信息自动添加到原始查询式或者翻译结果中

[26]，也可通过查询词建议的形式由用户选择相关词进行更为精确的查询扩展。在向查询式或者翻译结果中添加扩展词汇时，可以通过各种统计方法计算添加的词汇与查询式或者翻译结果的关联度，通过为不同扩展词设定不同权重的方式来减少某些弱相关词汇对检索结果的影响[27]。

跨语言查询扩展受到诸如翻译资源类型、翻译资源质量、语种等诸多因素的影响，在不同的实验中，查询扩展对最终检索结果的提高的效果不一致[28]，因此在查询扩展中需要慎重添加扩展词。作为一个辅助工具资源主题词表可以用来对传统方法获取的查询扩展词进行过滤。文献[29]利用目标语语料库进行翻译后的查询扩展时，利用双语词典将得到的用于查询扩展的目标语词汇转换为源语言，然后将其与源语言查询式在WordNet中的定义进行比较，从而剔除不相关的扩展词。Fredric C. Gey[26]等人基于传统的伪相关反馈策略，利用英语查询式搜索英文文档集合并从中提取前30篇相关文档的标引词，按标引词出现的文档数选取并添加到翻译后的查询式中。

上文介绍了多语主题词表在自动查询翻译和跨语言查询扩展中的应用情况，但是，总体来说，多语主题词表主要用于交互式的跨语言信息检索中，辅助用户进行查询式的构建和翻译、主题词表的浏览、查询扩展等，并出现了很多可嵌入到现有信息检索系统的交互界面[30,31]。基于多语主题词表的信息检索效率高，但是由于主题词表不能及时反映新事物的发展，概念数量有限、结构复杂，不易为非专业人员掌握，所以在实际的检索系统中很少为普通的用户所使用，消耗大量的人力物力对信息进行的标引和组织的结果在检索过程中的价值没有得到体现，这是一种巨大的浪费。未来研究人员应该更多考虑如何辅助普通用户更好利用多语主题词表，更好地满足多语言信息需求。

5 结语

随着不同语种网络信息的迅速增加，将数量庞大的资源关联起来并为用户提供服务成为多语言信息服务亟待解决的问题。多语主题词表以词汇规范控制为基础，采用概念和概念关系表示多语言知识的内在关联，语义颗粒度精细、规范，提供语义扩展机制，便于对资源进行语义层面的精细化描述和精确检索。国外关于多语主题词表的构建和应用等方面的研究比较重视，在理论和实践层面均有很好的成果，相比起来，国内对多语主题词表的研究重视程度不够。加强多语主题词表的相关技术研究，实现信息资源的语义化和多语言化标注，为用户提供具有丰富语义的、准确的跨语言信息资源，才能真正达到多语言信息服务的目标。

参考文献

[1] EuroVoc [EB/OL]. [2013-04-30]. http://eurovoc.europa.eu/drupal/.
[2] AgroVoc [EB/OL]. [2013-04-30]. http://aims.fao.org/standards/agrovoc/about.
[3] MeSH [EB/OL]. [2013-05-07]. http://www.nlm.nih.gov/mesh/.
[4] MeSH Browser [EB/OL]. [2013-05-07]. http://www.nlm.nih.gov/mesh/MBrowser.html.
[5] NELSON S J, SCHOPEN M, SAVAGE A G, et al. The MeSH Translation Maintenance System: structure, interface design, and implementation [C/OL]// Medinfo, San Francisco, America, September 7-11, 2004 [2013-03-05]. http://www.researchgate.net/publication/8353100_The_MeSH_translation_maintenance_system_structure_interface_design_and_implementation.
[6] ŠARI F, ŠNAJDER J, BAŠI B D, et al. Enhanced Thesaurus Terms Extraction for Document Indexing [C]// Proceedings of the 27th International Conference on Information Technology Interfaces, Cavtat, Hrvatska, July 20-23, 2005: 227-232.
[7] MARKÓ K, HAHN U, SCHULZ S, et al. Interlingual Indexing across Different Languages [C/OL]// RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, Avignon, France, April 26-28, 2004 [2013-04-06]. http://pdf.aminer.org/000/734/329/interlingual_indexing_across_different_languages.pdf.
[8] DAUDARAVICIUS V. The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance [C// 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010: 648-660.
[9] POULIQUEN B, STEINBERGER R, IGNAT C. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus [C]// Ontologies and Information Extraction. Workshop at EUROLAN’ 2003: The Semantic Web and Language Technology – Its Potential and Practicalities. Bucharest, Romania, July 28 – August 8, 2003: 9-19.
[10] MEDELYAN O, WITTEN I H. Domain-independent automatic keyphrase indexing with small training sets [J/OL]. Journal of the American Society for Information Science and Technology, 2008, 59(1): 1026-1040 [2013-03-05]. http://onlinelibrary.wiley.com/doi/10.1002/asi.20790/full.
[11] LAUSER B, HOTHO A. Automatic multi-label subject indexing in a multilingual environment [C/OL]// ECDL 2003, Trondheim, Norway, August 17-22, 2003:140-151 [2012-03-06]. http://www.kde.cs.uni-kassel.de/benz/hotho/pub/lauserhothoecdl03.pdf.
[12] JEX [EB/OL]. [2013-02-16]. http://ipsc.jrc.ec.europa.eu/index.php/Traineeships/60/0/.

[13] Agrotagger [EB/OL]. [2013-02-16]. http://aims.fao.org/agrotagger/.
[14] KEA [EB/OL]. [2013-02-16]. http://www.nzdl.org/Kea/.
[15] MEDELYAN O. Automatic Keyphrase Indexing with a Domain-Specific Thesaurus [D]. Breisgau, Germany: University of Freiburg, 2005.
[16] DE CAMPOS L M, FERNÁNDEZ-LUNA J M, HUETE J F, et al. Automatic Indexing from a Thesaurus Using Bayesian Networks: Application to the Classification of Parliamentary Initiatives [C/OL]// ECSQARU 2007, Hammamet, Tunisia, October 31 - November 2, 2007:865-877 [2013-04-10]. http://www.cs.rhul.ac.uk/home/aeromero/pdf/lncs07-ecsqaru-thesaurus.pdf.
[17] LANCASTER F W, WARNER A J. Intelligent Technologies in Library and Information Service Applications [M]. Medford, NJ, Information Today, 2001.
[18] CLEMENTS J. An Evaluation of Automatically Assigned Subject Metadata using Agrotagger and HIVE [R/OL]. [2013-04-10]. http://aims.fao.org/sites/default/files/files/Clements_FAO_Metadata_Assignment.pdf.
[19] SALTON G. Automatic Processing of Foreign Language Documents [J]. Journal of the American Society for Information Science, 1970, 21(3): 187-194.
[20] Eichmann D, RUIZ M E. Cross-Language Information Retrieval with the UMLS Metathesaurus [C]// Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, August 24-28, 1998: 72-80.
[21] GONZALO J, VERDEJO F, PETERS C, et al. Applying EuroWordNet to cross-language text retrieval [J]. Computers and the Humanities, 1998, 32(2/3): 185-207.
[22] YAROWSKY D. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora [C]// Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, Nantes, France, August 23-28, 1992: 454-460.
[23] HASNAH A M, JAAM J M. Thesaurus-based query disambiguation method for cross-language information retrieval [J]. International Journal of Intelligent and Cooperative Information Systems, 2002, 2(2):58-68.
[24] VOLK M, RIPPLINGER B, VINTAR S, et al. Semantic annotation for concept-based cross-language medical information retrieval [J]. International Journal of Medical Informatics, 2002, 67(1): 97-112.
[25] DÉJEAN H, GAUSSIER E, RENDERS J-M, et al. Automatic Processing of Multilingual Medical Terminology: Applications to Thesaurus Enrichment and Cross-Language Information Retrieval [J]. Artificial Intelligence in Medicine, 2005, 33(2): 111-124.
[26] GEY F C, JIANG H. English-german cross-language retrieval for the girt collection-exploiting a multilingual thesaurus [R]. CALIFORNIA UNIV BERKELEY DATA ARCHIVE AND TECHNICAL ASSISTANCE, 2006.
[27] SADAT F, YOSHIKAWA M, UEMURA S. Exploiting Thesauri and Hierarchical Categories in Cross-Language Information Retrieval [C/OL]// 5th International Conference, TSD 2002, Brno, Czech Republic, September 9-12, 2002:139-146 [2013-04-10]. http://www.fi.muni.cz/ tsd2002/papers/94_Fatiha_SADAT.pdf.
[28] GEY F, CHEN A. TREC-9 Cross-Language Information Retrieval (English – Chinese) Overview [C/OL]// Proceedings of the Ninth Text Retrieval Conference (TREC-9), 2001:15-23 [2013-04-10]. http://trec.nist.gov/pubs/trec9/papers/trec9-clir-overview.pdf.
[29] BELLAACHIA A, AMOR-TIJANI G. Enhanced Query Expansion in English-Arabic CLIR [C]// DEXA '08, Turin, Italy, September 1-5, 2008: 61-66.
[30] STAFFORD A, SHIRI A, RUECKER S, et al. Searchling: User-Centered Evaluation of a Visual Thesaurus-Enhanced Interface for Bilingual Digital Libraries [C]// Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries. Aarhus, Denmark, September 14-19, 2008: 117-121.
[31] SHIRI A, RUECKER S, ANVIK K. Thesaurus-enhanced visual interfaces for multilingual information retrieval [J]. Proceedings of the American Society for Information Science and Technology, 2006, 43(1): 1-7.

作者简介

徐红姣，硕士，信息管理专业，研究方向：跨语言信息检索。E-mail: xuhj@istic.ac.cn
张均胜，博士，计算机软件与理论专业，研究方向：多语言信息服务、语义计算。E-mail: zhangjs@istic.ac.cn
王惠临，研究员，博士生导师，研究方向：多语言信息服务、机器翻译、自然语言处理。E-mail: wanghl@istic.ac.cn

Research on Multilingual Thesaurus and Its Application

Xu Hongjiao, Zhang Junsheng, Wang Huilin / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: It's an important problem for multilingual information service to describe multilingual information precisely on semantic level and to provide accurate multilingual information resources. Multilingual thesaurus is one of the most effective tools to solve this problem. In this paper, three common used multilingual thesauruses are introduced firstly and then the problem about how multilingual thesaurus can be used in multilingual automatic indexing and multilingual information retrieval is analyzed. All of these show that multilingual thesaurus is a valuable tool in the area of multilingual information service.

Keywords: Multilingual thesaurus, Multilingual automatic indexing, Multilingual information retrieval

（收稿日期：2013-06-25）