DOI:10.16353/j.cnki.1000-7490.2010.11.022

● 高影繁，王惠临，徐红姣（中国科学技术信息研究所，北京 100038）

跨语言文本分类技术研究进展*

摘 要：本文以综述的形式对跨语言文本分类技术目前的发展态势进行了介绍，从应用背景出发，了解跨语言文本分类技术的社会需求；从关键技术出发，了解该项技术的核心问题及解决方案；从已有研究成果得到的结论揭示了该项技术的发展状况，作为一种重要的多语信息组织手段，跨语言文本分类技术发展前景广阔。

关键词：跨语言文本分类；特征提取；算法

Abstract: The present development situation of Cross-Language Text Categorization (CLTC) technologies is summarized. The paper describes the social demand for CLTC technologies from the perspective of the application background, describes the core issues of and solutions to CLTC technologies from the perspective of key technologies, and discloses the development status of CLTC technologies from the conclusions drawn from the obtained research results. As an important means for multilingual information organization, CLTC technologies have a broad development prospect.

Keywords: cross-language text categorization; feature extraction; algorithm

1 研究历程

1.1 应用背景

网络技术发展的日新月异促使多语种网络信息资源正在以惊人的速度不断丰富、增长。全球标准互联网用户调查和分析权威机构 Nielsen-NetRating 的调查数据显示，从 2000 年到 2007 年的 8 年间，全世界各种语言的网络使用增长率达到 244.7%。我国的多语种信息资源同样面临迅速膨胀的问题，以国家科技图书文献中心、国家科技数字图书馆为例，截至本文撰稿时为止，英文库期刊、会议、学位论文、科技报告等的馆藏数量达到 15 093 544 篇，俄日文库的期刊数分别为 163 753（俄文）和 872 790（日文）篇，中文库期刊、会议、学位论文多达 24 555 169 篇。毋庸置疑，这些多语种信息资源的数量将会持续快速地增长。

文本分类（Text Categorization，TC）作为一种重要的信息组织手段，其目的是将文本自动分入到已知类别中。到目前为止，文本分类算法的研究已经非常成熟了（如 KNN，SVM，Naive Bayes 算法等），基本进入到可以实用的阶段。现在的问题是，当机构或组织的文档管理部门日益依赖自动文本分类时，如何解决多语种文档的归类问题？现在很多搜索引擎都对 Web 页面进行层次分类以降低搜索的复杂性并提高精度，但如何把这种层次分类和语种结合起来呢？也就是说，在文档集中多于一种语言文档存在的情况下，怎样用同样的分类树分类这些异种语言的文档。

另外一个问题是多语主题检测与跟踪（Topic Detection and Tracking，TDT），它的任务是：监控各种语言的信息源，在新主题出现时发出警告，在信息安全、金融证券、行业调研等领域都有广阔的应用。多语 TDT 本来是跨语言信息检索技术的一个应用领域，但是当需要检测与跟踪的主题具有更高的抽象层次时，比如，需要监控“春季感冒”这个主题，跟踪各种语言的、与该主题相关的信息。很显然，这个主题包括“H1N1”、“普通感冒”、“病毒性感冒”等多个具体的主题，此时采用跨语言检索技术有些力不从心。

从以上的叙述中可以看出，在多语种信息高度膨胀的背景下，如何实现不同类别体系下多语种资源的高效组织和融合，将是多语信息组织领域亟待解决的问题。跨语言文本分类技术的出现为问题的解决提供了一条新途径。

1.2 相关研究成果

跨语言文本分类（Cross-Language Text Categorization，CLTC）研究起步较晚，最早的研究成果可以追溯到 1999

* 本文为“十一五”国家科技支撑计划项目（项目编号：2006BAH03B02），博士后科学基金（项目编号：20090450465）和中国科学技术信息研究所预研基金（项目编号：YY-200908）资金支持。

年。为了将主题检测和跟踪研究从英语扩展到中文，研究者给出几篇种子文档，并用它们来监测信息流中具有相同主题的文档[1]。研究结果表明，采用由中文数据训练的分类器分类英文文档时，效果要比单语言分类器差很多。跨语言文本分类作为一个学术概念被明确提出是在2003年，N. Bel等人指出：所谓的跨语言文本分类，是指在不需人工干预的情况下将现有的自动文本分类系统由单语言扩展到两种或多种语言，并在英语—西班牙语双语语料上进行了文本分类实验[2]。L. Rigutini，A. Gliozzo，J. S. Olsson，Y. Y. Li以及B. M. Amine等人分别在2005—2007年间年研究了英语—意大利语、英语—捷克语、英语—日语、英语—西班牙语的跨语言文本分类问题[3-7]。英语—中文跨语言分类的研究成果直到2008年才见诸公开的学术文献，美国马里兰大学的Y. J. Wu和上海交通大学的K. Wu、L. Xiao等人分别在重要国际会议上发表了他们的中—英跨语言文本分类的研究成果[8-10]。

因为这是一个新兴的研究领域，目前能够找到的跨语言文本分类文献非常少，且研究都集中在实验室阶段。即便如此，还是可以从文献中找到它的技术脉络，从而分析其中的关键环节，为下一步的实际应用打下重要的基础。

2 关键技术

跨语言文本分类与跨语言信息检索技术师出同门、异曲同工。跨语言检索技术虽然离不开检索，但它的关键点却是在翻译的环节。即，如何将源语言查询和目标语言文档转换到同一个语言空间，查询翻译技术、翻译消歧技术是核心。跨语言文本分类技术也是这样，虽然离不开分类，但必须面对源语言文档训练集与目标语言待分类文档的空间转换问题。与语言空间变换相结合的跨语言文本分类算法是核心环节，这其中又包含两种情况：一是基于跨语言文本特征提取的常规文本分类算法；二是结合跨语言文本特征提取的新算法；本文依据不同的语言空间转换手段阐述跨语言文本分类的关键技术。

2.1 基于平行语料的跨语言分类方法

Y. Y. Li和A. Gliozzo等采用KCCA（Kernel Canonical Correlation Analysis）、LSI（Latent Semantic Indexing）等方法，不依赖于翻译资源，以平行（可比）语料为基础完成源和目标语言的空间转换。LSI方法是跨语言信息检索领域常用的仅依据平行语料就可以进行源和目标语言空间转换的方法。同样以平行语料为基础，Y. Y. Li采用以寻找两种语言的文档间最大关联关系为目标的KCCA方法，作者先将该方法用于“日—英”跨语言检索，实验结果优于LSI方法，但计算复杂度较高；再将KCCA方法用于日、英两种语言间的语义对应，结合SVM算法完成“日—英”跨语言文本分类任务。

A. Gliozzo的实验中采用了更容易获得的可比语料。在没有任何外部资源（如：双语词典等）而仅仅依靠可比语料的情况下，采用非监督方法从平行语料中获取多语领域模型MDM（Multilingual Domain Models）（注：该方法基于LSI），该模型被用来在多语文档间定义泛化的相似性函数（如核函数），继而被应用于SVM中完成“英语—意大利语”的跨语言文本分类任务。值得注意的是，作者假设在不同语言中同样的实体用相同的词表示，允许通过观察词形自动检测翻译等价对，并假定在可比语料中包含大量这样的词。该假设限制方法的应用范围，不适用于英—日、英—中、英—韩等词形差异比较大的情况。

2.2 基于机器翻译软件的跨语言分类方法

L. Xiao等认为源语言和目标语言的空间转换过程可能会导致以下问题：①由于语言和文化的差异，在空间变化时会发生主题迁移（Topic Drift）。②翻译过程中会产生翻译错误。③源语言和目标语言文档的特征空间可能不同。主题迁移和翻译错误会导致源和目标语言数据分布存在差异，这违反了大部分现有文本分类算法有关训练集和测试集数据分布一致性的基本假设。为解决这个问题，作者提出基于信息瓶颈（Information Bottleneck，IB）理论的跨语言文本分类算法，采用成熟的商业机器翻译软件Google Translator完成“中—英”的语言空间转换任务，算法的核心是利用两种语言数据集中的公共部分（Common Part）来提高分类器性能。与Naive Bayes，SVM，TSVM等算法的比较结果验证了作者算法的高效性。

与文献［10］相比，B. M. Amine等的方法更加简单易行。作者采用能提供14种语言翻译功能的JWT（Java Web Translator）软件完成“英文—西班牙文”的文档翻译工作，并利用WordNet生成文档的概念空间（作者认为WordNet中包含的上下位关系等知识对纠正翻译错误有很大帮助）。在这些工作基础上，通过待分类文档的概念向量与各个类别概念向量的相似度值判断文档所属的类别。

L. Rigutini等研究在文档数量不均衡情况下的跨语言文本分类问题。这种情况在现实中非常普遍，即存在一种语言的少量有类别标注的文档，和另一种语言的大量没有类别标注的文档，以及如何对这些未标注文档进行分类的问题。作者采用EM算法解决“英语—意大利语”的源和目标语言的主题迁移问题：首先采用Office Translator Idiomax 2005翻译软件将有类别标注的源语言文档转换为目标语言，并采用生成的目标语言文档训练分类器（作者实验采用Naive Bayes算法）；然后，利用新生成的分类器对未标注的目标语言文档进行分类，并依据EM算法逐渐调整分类器参数，直至归于稳定。作者通过引入目标语言文档

词汇来解决源和目标主题偏移问题的思路非常值得肯定。

2.3 基于词典的跨语言分类方法

这是最常用跨语言空间转换手段。相比较前两种翻译资源，双语词典更容易获得且成本较低。N. Bel 等通过构建“英语—西班牙语”的领域术语翻译表，进行术语翻译（Terminology Translation）和类别文档中已有的词汇翻译（Profile-Based Translation），并据此生成文档向量以备 Rocchio 和 Winnow 分类算法使用。作者认为，仅翻译类别中出现的词汇是一种代价低廉的跨语言分类方法（作者实验中每个类别平均翻译 60 个词），虽然分类精度稍微有点低，但性价比较好，方法趋于实用。

J. S. Olsson 的算法基于向量空间模型，在已有英语训练文档的前提下，作者巧妙地利用捷克语到英语的翻译概率矩阵和捷克语文档向量的乘积，得到捷克语文档的英语向量表示形式，继而可以利用文本分类算法对待分类的捷克语文档进行分类（作者采用了 KNN 算法）。作为翻译资源的翻译概率词典主要用于构建翻译概率矩阵。实验结果表明，作者的跨语言系统可以达到单语言系统 73% 的分类精度。

K. Wu 的方法与 L. Rigutini 的方法颇为类似。不同的是，前者采用的翻译资源是双语概率词典而后者采用机器翻译软件。为保证翻译的准确性，作者利用双语电子词典和手边的训练语料集，应用 EM 算法导出给定类别下翻译的条件概率。作者将分类的过程分为两个阶段，第一阶段，根据翻译为目标语言的源语言文档建立分类器（Naive Bayes），赋予目标语言文档初始类别标记；第二阶段，引入 EM 算法对目标语言类别标记进行修正，获得文档最终的类别标记。

Y. J. Wu 等所做的工作更确切地说是双语主题的分面分类（Bilingual Topic Aspect Classification）。先将文档细致划分为一个个子主题（SubTopic），然后进行跨语言分类。首先采用双语概率词典进行语言空间转换；其次因为子主题下的文字内容较短，作者因此可以利用 LSA 等技术进行特征的维数约减；最后采用基于向量权重表示的改进的 KNN 算法进行分类操作。作者的实验结果表明，在对某种语言文档进行分类时，分类器的训练应以该种语言文档为主，异种语言文档只能作为补充，如果数量过多的话会造成分类器性能的下降。

3 结论与展望

根据跨语言文本分类技术的已有成果和上文对关键技术的分析，可以得出如下结论：

1) 主题迁移是跨语言文本分类的主要壁垒。跨语言文本分类与单语言文本分类最大的差别在于训练和测试文档分属不同的语种，而文本分类算法的基本假设是训练集和测试集数据分布的一致性。在这种情况下，将异语种的数据加入源语言的训练文档集成为主要的解决方法，L. Xiao，L. Rigutini，K. Wu，Y. J. Wu 等都在实验中进行了尝试，结果说明了方法的可行性。基于平行（可比）语料的方法能够有效避免主题迁移问题，但语料的建设成本较高。

2) 重特征提取、轻分类算法。在单语言文本分类技术中，分类算法的研究已经相当成熟了，基本进入实用阶段。在跨语言文本分类中也是如此，研究者基本上都把重点放在了跨语言特征提取上，而其中的核心环节就是语言空间转换。基于平行语料、机器翻译软件和词典是 3 种常用的语言空间手段，前两者准确性高门槛也高（平行语料较难获得，好的商业机器翻译系统使用受限），后者获得成本低更趋于实用。从特征提取与分类算法的结合程度上来看，大部分情况是两者分离的，如文献［2，4－8］等；而文献［3，9－10］等利用目标语言特征调整源语言的分类器参数，是特征提取和算法结合的实例。

3) 多翻译资源的共同使用。无论在跨语言检索还是跨语言文本分类领域，在同一系统中使用多种翻译资源都是一种趋势，已有的跨语言文本分类研究成果印证了这一结论。比如，B. M. Amine 在使用翻译软件 JWT 的基础上，结合 Wordnet 词典生成概念向量；A. Gliozzo 等为使其方法扩展到任意语种间（文献［4］的方法仅用于词形相似的两种语言），在方案中加入了 MultiWordNet 和 Collins 词典资源[11]；文献［9］的研究中采用了成本较低的双语词典，为了提高翻译的准确性，在算法中加入了带类别信息的训练语料等。

4) 研究停留在实验室阶段。跨语言文本分类技术研究起步较晚，到目前为止仅有十几篇公开发表的研究成果，可以说研究处于非常初级的阶段。在这个时期，学者们更注重研究方法的可行性，为了提高分类精度，采用了很多计算成本较高的方法，如 LSI，KCCA，EM，基于信息瓶颈理论的方法等；在采用商业机器翻译系统时并没有考虑到算法的扩展性；对跨语言文本分类语料的建设非常不成熟，没有公开的测试集。跨语言文本分类的研究要达到实用水平还有很长的路要走。

对跨语言文本分类技术来说，虽然存在起步晚、研究成本较高、研究点较为分散等若干缺点，但依然可以看到其广阔的发展空间。面对日益增长的多语种科技信息资源，集成了跨语言文本分类技术的多语信息组织方法将在很大程度上提高多语资源的管理效率，为进一步拓展信息资源的多语言服务能力打下良好的基础。□

（下转第 104 页）

自己的借阅信息，也可以方便地查询所需新老馆文献的全部书目信息，而且还能方便、准确地找到新馆图书文献所在书架的位置信息。在建立书架 RFID 标识的同时建立了相关的文献查询定位导航系统，在各楼层、分区、巷道之间设立直观的指引和位置指示图，并在 OPAC 查询的结果页面中建立文献的具体位置指示和导向系统，让读者可以依据新图书馆书库内的参照物对比自己所处的位置，快速地寻找到所需要的图书。

4 结束语

该设计方案适应新馆和老馆两种不同文献管理模式并存条件下，对新馆和老馆文献通借通还的流通管理需要，同时也保证了与老文献管理系统的平滑对接，使原一卡通用户既可以享受新馆 RFID 智能管理系统带来的文献流通管理的便利，又可以通畅地利用老馆图书文献资源，没有给用户增加额外的负担。随着 RFID 技术的不断完善和硬件设备成本的降低，RFID 图书馆智能管理系统的开发成本也会继续降低。相信在不久的将来，RFID 图书馆智能管理系统会得到越来越广泛的应用，图书馆的文献管理也将更加人性化和智能化。□

参考文献

[1] 江先斌 . RFID 在现代图书馆中的应用研究 [J]. 海峡科学，2007 (9)：16-17.

[2] 张猛 . RFID 技术在图书馆的应用与发展 [J]. 计算机与网络，2009，35 (3)：138-141.

[3] 康东，石喜勤，等 . 射频识别核心技术与典型应用开发案例 [M]. 北京：人民邮电出版社，2008.

[4] 冯波 . 基于 RFID 的图书馆管理系统的研究 [D]. 武汉：武汉理工大学，2008.

[5] 甘琳 . dILAS 与 RFID 在深圳图书馆新馆的应用 [EB/OL]. [2009-10-25]. http://www.szln.gov.cn/UserFiles/contribute/1210041834505.doc.

[6] 江丽，伍萍 . 在 ILASⅡ2．0 图书馆管理系统下构建 RFID 系统——以武汉图书馆为例 [J]. 图书馆界，2008 (4)：39-42，48.

[7] 刘金华 . 总分馆体制下分步实施 RFID 的探讨[EB/OL]. [2009-10-20]. http://www.lib.stu.edu.cn/accessory/RFID9.pdf.

[8] 上海阿法迪智能标签系统技术有限公司 . RFID 技术在集美大学图书馆中的应用 [J]. 中国自动识别技术，2006 (2)：56-57.

[9] 刘白秋 . 无线射频识别技术在国内图书馆中的首次应用实践 [J]. 图书馆学研究，2007 (4)：10-12.

[10] 刘长征，熊璋，王剑昆 . 基于智能标签的射频识别系统的研究和实现[J]. 计算机工程，2003，29(20)：162-164.

[11] 董京曦 . 图书馆射频识别系统选型综合问题分析 [J]. 现代图书情报技术，2006 (S)：110-113.

[12] 李秀霞 . VTLS RFID 图书馆管理系统的结构与实现 [J]. 图书馆学研究，2009 (3)：27-29.

作者简介：欧朝静，女，1975 年生，馆员，硕士。

收稿日期：2010－08－02

(上接第 128 页)

参考文献

[1] WAYNE，C. Topic detection and tracking in English and Chinese [C]. IRAL，Hong Kong，2000：165-172.

[2] BEL N，et. al. Cross-lingual text categorization [C] // Proceedings of 7th European Conference on Research and Advanced Technology for Digital Libraries，2003：126-139.

[3] RIGUTINI L，et. al. An EM based training algorithm for cross-language text classification [C]//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence，2005.

[4] GLIOZZO A，STRAPPARAVA C. Cross language text categorization by acquiring multilingual domain models from comparable corpora [C] //Proceedings of the ACL Wordshop on Building and Using Parallel Texts. 2005：9-16.

[5] OLSSON J S，OARD D，HAJIC J. Cross-language text classification [C] // Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval，2005：645-646.

[6] LI Y Y，SHAWE-TAYLOR J. Using KCCA for Japanese-English cross-language information retrieval and document classification [J]. Journal of intelligent information systems，2006，27 (2)：117-133.

[7] AMINE B M，MIMOUN M. WordNet based multilingual text categorization [C]. 2007 IEEE/ACS International Conference on Computer Systems and Applications，Amman，Jordan，2007：13-16.

[8] WU Y，OARD W. Bilingual topic aspect classification with a few training examples [C] //Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，2008：203-210.

[9] WU K，LU Bao-liang. A refinement framework for cross language text categorization [C] //Proceedings of 4th Asia Information Retrieval Symposium，2008：401-411.

[10] XIAO L，et al. Can Chinese Web pages be classified with English data source? [C] //Proceeding of the 17th International Conference on World Wide Web，2008：969-978.

[11] GLIOZZO A，STRAPPARAVA C. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization [C] // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL，2006：553-560.

作者简介：高影繁，女，1974 年生，博士。
王惠临，男，研究员，博士生导师。
徐红姣，硕士。

收稿日期：2010－06－18