跨语言检索中机器翻译技术的应用和进展

□刘群 骆卫华 / 中国科学院计算技术研究所 北京 100080

摘要：本文以跨语言检索为背景，主要介绍了机器翻译技术的应用和进展。文章介绍了跨语言检索中机器翻译技术的应用形式，简单回顾了机器翻译技术发展历史中出现的各种方法及其基本思想和优缺点、特别是近年来统计机器翻译的发展。本文通过一个实例较为详细地介绍了目前主流的基于短语的统计机器翻译原理，然后通过一个实际的新闻长句子对几个典型的基于规则的和基于统计的机器翻译系统的翻译结果给出了细致的比较和分析。最后对机器翻译在跨语言检索中的应用前景做出了展望。

关键词：跨语言检索 统计机器翻译

1 引言

跨语言信息检索问题研究的是基于一种自然语言构造的、查询搜索任意语言文档的方法。它与信息检索、机器翻译两项技术有着天然的紧密联系。

很多跨语言检索方法都要用到不同形式的翻译技术。通常比较简单的方法就是直接查词典。因此词典构造和查询方法就成为这一类方法的重要研究内容。也有一些跨语言检索方法直接利用机器翻译系统进行翻译。

具体来说，跨语言检索中用到机器翻译的地方主要有两方面：

(1)查询语句的翻译。这是最普遍的用法。也就是将源语言的查询语句翻译成目标语言，再到目标语言的语料库中去查找相关的文本。

(2)目标文本的翻译。由于查询者可能只懂源语言，所以对于查询到的目标语言文本，需要翻译成源语言。这种翻译可能是全文翻译，也可能是对文本的摘要进行翻译。理论上说，我们也可以在检索之前，就将所有的目标语言文本都翻译成源语言，直接用源语言的查询语句进行检索，这样就不需要进行查询语句的翻译了。当然，由于这样做代价太高，现在实际上很少有人这么做。

由于效率原因，目前主流的做法还是对查询进行翻译。显而易见，在跨语言检索中，查询翻译结果的好坏，实际上对检索的效果起到了非常重要的作用。对于查询翻译，目前大部分系统采用的做法还是词典查询，然而随着近年来统计机器翻译技术的迅速发展，将统计机器翻译的研究成果应用于跨语言检索必将引起人们更多的重视。

近年来，国际上机器翻译研究取得了重大的突破。在一些主要的国际机器翻译评测中，采用传统的基于规则的机器翻译方法的系统的性能已被目前主流的统计机器翻译方法全面超越。其中的很多思想和做法已经被跨语言检索研究引入，为该领域的研究提供了新的思路。本文将主要从机器翻译的角度介绍这个领域的发展，重点是统计机器翻译近年来取得的进展，其中穿插了这些技术在跨语言检索中的应用。

本文将首先介绍机器翻译研究的历史和现状，以及目前主流的基于短语的统计机器翻译方法的基本原理，再通过一些实例，具体比较两个机器翻译系统——一个是传统的基于规则的机器翻译系统，另一个是基于短语的统计机器翻译系统的翻译结果，并分析其对跨语言检索可能产生的影响，最后给出结论和展望。

2 机器翻译研究的历史和现状[1]

机器翻译的历史，可以追溯到1946第一台现代电子计算机ENIAC诞生后不久。英国工程师布斯（A.D.Booth）和美国洛克菲勒基金会副总裁韦弗（W.Weaver）在讨论电子计算机的应用范围时，就提出了利用计算机进行语言自动翻译的想法。1949年，

[1] 关于机器翻译的历史更详细的介绍，请参见参考文献[1]和[2]。本文有关机器翻译历史的很多介绍材料都来自冯志伟先生的相关论著。

韦弗发表了一份以《翻译》为题的备忘录，正式提出了机器翻译问题。1954 年，美国乔治敦大学在国际商用机器公司（IBM 公司）的协同下，用 IBM-701 计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。在世界范围内，大量的资金和研究人员都投入到了机器翻译的研究之中。我国也是世界上最早开展机器翻译研究的国家之一，1956 年，国家便把机器翻译研究列入了我国科学工作的发展规划，成为其中的一个课题。1957，中国科学院语言研究所与计算技术研究所合作，开展俄汉机器翻译试验，翻译了 9 个不同类型的、较为复杂的句子。

这时候的机器翻译系统通常都是很简单的，大多采取了词典查询和简单词序调整的方法，这种方法被称为直接翻译方法。显然，这时候的研究者还没有意识到机器翻译的难度，这种简单的直接翻译方法对于个别的例子还可以凑效，但对于稍微复杂一些的句子就无法正确翻译了。实际上，这次机器翻译的研究热潮很快就碰了壁。

1964 年，为了对机器翻译的研究进展作出评价，美国科学院成立语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称 ALPAC 委员会），调查机器翻译的研究情况，并于 1966 年 11 月公布了一个题为《语言与机器》的报告，简称 ALPAC 报告，对机器翻译采取否定的态度。报告宣称："在目前给机器翻译以大力支持还没有多少理由"；报告还指出，机器翻译研究遇到了难以克服的"语义障碍"（semantic barrier）。应该说，这个评价还是基本客观的。不过这个报告对整个机器翻译的研究产生了非常消极的影响，世界范围内的机器翻译研究热潮一下子就消退了，很多政府都撤销了这方面的研究经费，使得机器翻译的研究进入了萧条期。

虽然如此，很多执着的研究者并没有放弃对机器翻译的研究，一些国家也依然在一定程度上支持这方面的工作。这时候，人们逐渐认识到，要完成机器翻译，计算机必须在一定程度上"理解"源语言的句子。与此同时，人工智能这门学科在 1970 年代也有了很大的发展，各种知识表示和知识推理的理论和算法被研究者提出来。在这种背景下，人们对自然语言理解和机器翻译的认识更为深刻了。

这一阶段机器翻译方法的主要特点是对语言进行了深层次的分析、转换和生成。也就是说，翻译不再是在句子的表层（词序列）上进行，而是在句子的某种更深层结构（如句法结构、语义结构或知识表示）的层面上进行。为了做到这一点，需要大量的语言知识和翻译知识，因此，这时候的机器翻译程序，从结构上比早期机器翻译程序的一大进步是采用了数据与程序相分离的形式。语言知识和翻译知识以数据形式存在，而翻译程序利用这些数据进行翻译。这种数据，最常见的表现形式就是规则和词典。因此，这一类机器翻译方法被称为基于规则的机器翻译方法。

机器翻译方法也可以根据源语言理解和翻译转换所在的语言学层面的不同进行划分。如图 1 所示。

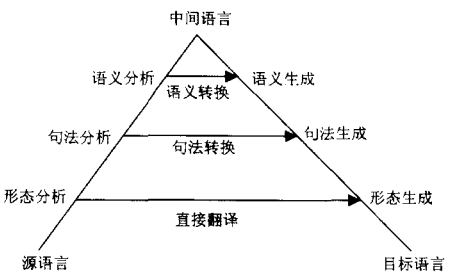

图 1　机器翻译方法图示

在词层面进行翻译的方法成为直接翻译法，早期的机器翻译方法就是这一类。基于规则的机器翻译方法通常在更深的层次上进行机器翻译，包括句法层面（句法转换）、语义层面（语义转换），甚至知识层面（中间语言）。

基于规则的机器翻译方法导致了程序工作者和语言工作者的一种合作范式：程序工作者和语言工作者共同确定翻译算法、制定数据规范。确定语言知识和翻译知识的表示形式；然后程序工作者编写程序实现翻译算法，这种翻译算法的运行是用语言知识和翻译知识来驱动的；而语言工作者编写语言知识和翻译知识。

在这种工作范式下面，系统翻译性能通常受到两方面因素的制约：一是算法的设计是否合理，另一个是语言知识是否足够丰富。其中最主要的瓶颈还是后者。一旦翻译程序编程结束，并经过调试稳

定以后，基本上就不需要再做修改了。而改进翻译性能的任务完全落在了语言工作者的身上。对于基于规则的机器翻译系统而言，知识获取实际上是最大的瓶颈。通常，经过一个人一年左右的调试，就很容易得到一个可以翻译简单句子的演示系统。但要得到一个真正初步实用的机器翻译系统，非得要通过一批人常年累月的调试和积累不可。

基于规则的机器翻译系统，在1980年代，达到一个高峰期。市场上出现了很多的机器翻译系统，其中一些系统进入了初步的实用化阶段。国际上也出现了一些大规模的机器翻译系统研究计划，如欧盟的Eurotran项目和日本的亚洲五国语言机器翻译项目。与此同时，基于规则的机器翻译系统的问题也逐渐暴露出来。这类方法最大的问题就是知识获取问题。依靠语言工作者人工编写规则，似乎永远也不能满足实际应用的需要。一个在市场上销售的机器翻译系统，通常都要经过数十乃至数百人年的调试，但翻译质量还是远远不能达到令人满意的程度。更为糟糕的是，人工添加规则的做法，导致规则库的规模更大，系统性能的改进就更困难。因为一方面，规则库越大，规则之间的冲突就越多，导致所谓的“跷跷板现象”，系统虽然对某些句子翻译效果好了，但对另外一些句子效果反而差了，系统的整体性能并没有提高；另一方面，越到后面加入的规则，通常都是一些粒度非常小的规则，只是处理非常个别的语言现象，对系统整体性能的改进很小，整个系统的性能提高极为缓慢。而实际上，这个时期一些大型的机器翻译研究项目也都以失败而告终。基于规则的机器翻译方法似乎走到了尽头。

同一时期，信息检索技术的先驱者也已经开始了跨语言检索的研究。1960年代，SMART系统的开发者Salton就在小规模数据集上进行了跨语言检索的实验。那时使用的方法也是相对简单粗糙的，基本上就是通过查词典把查询和文档统一到同一种语言。为了克服查词典这种方法的两个根本缺陷：覆盖率有限和词汇歧义，研究者开始追求构造越来越大的词典，并不断向其中增加人工和统计知识。但是，这种方法永远赶不上语言变化的速度，而构造词典又是一项代价高昂的工作，因此人们渐渐放弃单一的查词典方法，代之以查词典与其它方法相结合的策略。

为了克服基于规则的机器翻译方法所面临的困难，从1980年代末到1990年代初，出现了一类新的机器翻译方法，可以统称为基于语料库的机器翻译方法。其基本思想是，采用语料库作为机器翻译知识的来源，直接利用各种算法，从语料库中获取翻译知识，而不需要依靠人工方法来编写翻译知识。研究者希望，这一类方法能够克服传统方法中的知识获取瓶颈问题。具体来说，基于语料库的机器翻译方法又可以分为两种，一种是基于实例的机器翻译方法，一种是统计机器翻译方法。

基于实例的机器翻译方法采用类比推理的思想，在翻译一个句子的时候，到语料库中去寻找与被翻译句子相似的句子或者句子片断，通过对语料库中这些相似句子或者句子片断的修改和组合，得到输入句子的译文。统计机器翻译的基本思想是，认为源语言句子到目标语言句子的翻译是一个概率问题，任何一个目标语言句子都有可能是任何一个源语言句子的译文，只是概率不同，机器翻译的任务就是找到概率最大的句子。因此，统计机器翻译又可以分为一下几个问题：模型问题、训练问题、解码问题。所谓模型问题，就是为机器翻译建立概率模型，也就是要定义源语言句子到目标语言句子的翻译概率的计算方法。而训练问题，就是要利用语料库来得到这个模型的所有参数。所谓解码问题，就是在已知模型和参数的基础上，对于任何一个输入的源语言句子，去查找概率最大的译文。

基于实例的机器翻译方法和统计方法虽然都属于基于语料库的翻译方法，但它们的发展经历了一个非常不同的过程。基于实例的机器翻译，曾经受到了广泛的关注，研究的人很多。但由于各研究者采用的方法虽然总体上都是采用类别推理的思想，但实现方法上都不尽相同，没有形成一种被普遍接受的模式，因此总体上，基于实例的机器翻译研究比较分散，虽然这种机器翻译系统体现出了一些基于规则的机器翻译系统所没有的优点，但在总体性能上，并没有明显超出传统的规则方法。

统计机器翻译方法的发展也经历了一个曲折的过程，但现在已经成为机器翻译研究的主流[3]。统计机器翻译最早是由IBM公司的研究者在1990年前后提出来的，他们开发的系统在美国ARPA组织的机器翻译评测中取得了可以与Systran系统相媲美的结果。

要知道，Systran 系统经过了几十年的调试，而 IBM 的系统只经过了几年的开发，直接从语料库中获取翻译知识，没有经过人工的规则调试。IBM 的工作在当时引起了轰动。但由于当时 IBM 动用了最先进的工作站集群计算环境，其它研究者很难重复他们的工作，所以在很长一段时间内统计机器翻译一直停滞不前。一直到 1999 年，一些研究者在一次约翰霍普金斯夏季研讨会上，重复了 IBM 的工作，并且发布了一个开放源代码的工具以后，统计机器翻译重新引起了人们的重视。2002 年开始，美国国家标准技术研究所（NIST）在美国高级研究计划局（DARPA）的支持下，开展了一个每年一度的机器翻译评测工作[4]，在这个系列评测中，统计机器翻译方法一鸣惊人，全面超过了传统的基于规则的机器翻译方法，成为了目前机器翻译研究的主流。图 2 是近几年 NIST 评测中最好的系统的评测结果：

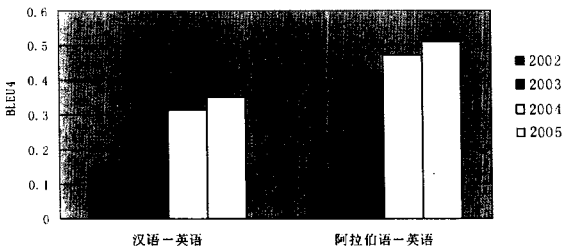

图 2 NIST 机器翻译测评结果比较

虽然由于每年的测试集都有所不同，各年的成绩并不可比，但实际上，每年的测试集的难度是大致相同甚至略有增加的，因此，我们可以看到，每年的机器翻译成绩实际上都有不小的提高，几乎是呈线性增长的趋势。这引起人们非常大的兴趣。因为大家都知道，基于规则的方法在系统达到一定水平后，再要提高一点点都是非常困难的。而统计机器翻译方法的水平却年年不断地在提高，这不禁使人对其充满了希望。这就导致近几年来形成了一个新的统计机器翻译研究的热潮，参与这个领域研究的人数和发表论文数量都呈指数型增长。图 3 是 Google 公司统计的统计机器翻译研究的论文数量统计，从中我们可以看到这种趋势。

统计机器翻译技术本身也经过了一个不断发展的过程。统计集群翻译的模型框架从早期的噪声信道模型发展到目前普遍采用的对数线性模型，其中

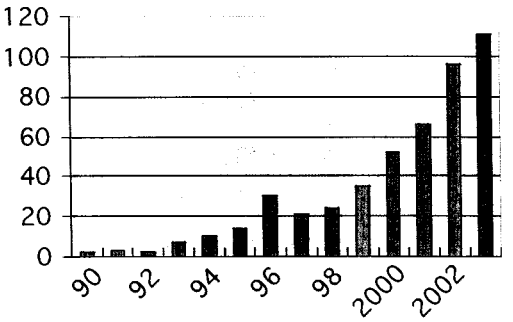

图 3 统计机器翻译研究的论文数量

最主要的统计翻译模型也从早期的基于词的模型发展到了目前主流的基于短语的模型，以及目前很多人都在进行基于句法统计翻译模型研究。

在统计机器翻译研究中，计算机工作者和语言工作者的合作形成了一种新的范式。语言工作者的工作主要是定义和开发语料库、词典等语言学资源，而计算机工作者主要是改进算法。这种范式比基于规则的机器翻译系统研究中形成的开发范式更加有效。一方面，语言数据和算法之间的区分更加清楚，在系统开发过程中，基本上不用语言工作者和计算机工作者进行交互，二者独自开发就可以了；另一方面，语言工作者开发的语言资源不是为特定的机器翻译系统服务的，可以用于任何一个机器翻译系统，这样从总体上大大减少了语言工作者的重复劳动，形成了良好的积累效应。而且，计算机工作者可以不断地通过改进算法来提高机器翻译系统的性能，而不像基于规则的范式中，一旦算法确定，提高系统性能的任务主要取决于语言工作者的词典编写和调试。

和机器翻译一样，语料库方法也成为跨语言检索研究者的新宠。对语料库的使用目前有两种不同的做法：一是把语料库作为自动构造词典的工具，二是把语料库作为查询翻译模型的训练数据。前者利用文档对齐、句子对齐乃至词对齐的语料抽取词的翻译知识，从而实现词典的增量式自动构造，再使用查词典技术来进行查询翻译和检索，这项技术仍未摆脱简单的查词典模型的窠臼，但由于语料中包括共现等有用的统计信息，通过这种方法实现的跨语言检索仍然取得了较大的性能提升。而后者则直接把统计信息用于查询翻译，避免了舍近求远的过程，通过词向量翻译模型或隐含语义标引，在一定程度上消除了翻译歧义问题。整体来看，跨语言

检索与机器翻译在使用语料库时面临着基本相同的问题，但也有一些地方有所不同。比如，有研究表明，跨语言检索对于句法分析产生的错误不如机器翻译敏感，但对于语义分析则要依赖得多。

3 基于短语的统计机器翻译原理

目前，最成熟的统计机器翻译方法还是基于短语的方法[5]。这一类方法本质上并不复杂。下面我们简单介绍一下这种方法的原理，以便读者对这一方法有个初步的了解。

基于短语的统计机器翻译方法，其基本的思想其实非常简单，就是从双语对齐的平行语料库中，抽取所有可能的双语词语序列，并记录其概率信息。对于新输入的源语言句子，首先查找语料库中所有已经存在的双语词语序列，取出其中的目标语言词语序列，并根据其概率，组合得到最优的译文句子。

我通过一个例子说明一下。

统计机器翻译分为训练和翻译（又成为解码）两个过程。

训练的时候，最初输入的是一个双语语料库，实际上就是一对一对互为翻译的句子。如下所示：

中国化工工业保持稳定增长。
China's chemical industry maintains steady growth.

万里会见泰国客人
Li Wan meets with guests from Thailand

……

经过一个词语对齐过程，得到如下形式的数据：

中国 化工 工业 保持 稳定 增长 。
China's chemical industry maintains steady growth .

万里 会见 泰国 客人
Li Wan meets with guests from Thailand

……

可以看到，我们已经知道了句子中哪些词是互为翻译的。下面，就可以进行短语抽取了。所谓短语抽取，就是抽取出语料库中所有互为翻译的连续的词串，而不用管这个词串是否具有真正的含义。以上面第二个句子为例，我们可以抽取到以下的短语：

万里 || Li Wan
万里 会见 || Li Wan meets with
万里 会见 泰国 客人 || Li Wan meet with guests from Thailand
会见 || meets with
会见 泰国 客人 || meet with guests from Thailand
泰国 || Thailand
泰国 客人 || guests from Thailand
客人 || guests

实际上，所抽取出来的短语中，除了两种语言的词语外，还有一些概率信息。下面是我们从真实的大规模语料库中抽取出的一些以“大规模”开头的短语片断。

中文短语	英文短语	概率
大 规模	Mass	0.468586
大 规模	Massive	0.227749
大 规模 的	large-scale	0.446154
大 规模 的	Massive	0.415385
大 规模 的 救济	large-scale relief	0.666667
大 规模 的 救济	largescale relief	0.166667
大 规模 的 救济 行动	large-scale relief operations	0.5
大 规模 的 救济 行动	massive relief operation	0.5
大 规模 军事	large-scale military	0.423077
大 规模 军事	largescale military	0.173077
大 规模 军事 行动	large-scale military action	0.290323
大 规模 军事 行动	large-scale military operation	0.193548
大 规模 杀	, while widespread	1
大 规模 杀伤	mass killing	1
大 规模 杀伤	large scale killings	0.038462
大 规模 杀伤	mass destruction	0.961538
大 规模 杀伤 武器	Lethal weapons of mass	0.041667
大 规模 杀伤 武器	weapons of mass destruction	0.958333
大 规模 杀伤性	mass destruction	0.821429
大 规模 杀伤性	mass destructive	0.178571
大 规模 杀伤性 武器	mass destructive weapons	0.181818
大 规模 杀伤性 武器	weapons of mass destruction	0.8

我们看到，“大规模”这个短语，在语料库中主要有mass、massive、large-scale这几种翻译方法，跟不同的词搭配的时候，有一些习惯性的翻译方法，通常不能混用。比如，指“军事行动”或“救济行动”的时候，通常用large-scale，指“救济行动”的

时候，也可以用massive，而指“杀伤性武器”的时候，通常翻译成mass。对于“行动”一词，“救济行动”一般翻译成relief operations，而“军事行动”可以翻译成military action，也可以翻译military operations。“大规模杀伤性武器”，决大部分情况下都是翻译成“weapons of mass destruction”，少数情况下翻译成“mass destructive weapons”。要知道，上述这些知识全部都是机器自动获得的，而如果在基于规则的方法中，要依靠人写的规则来区分如此细微的差别，是非常困难的。而对于跨语言信息检索的查询翻译来说，由于输入的查询语句通常都是一些短语片段，采用这种基于短语的翻译模型无疑是非常合适的。

从上面的介绍中，读者也会发现很多问题。最大的问题就是，统计机器翻译似乎回到了早期不需要理解的时代，是在词语或短语层面进行翻译，而没有在更深层的句法或者语义层面进行转换。事情确实如此，很多人尝试在统计机器翻译方法中引入句法分析等方法，但效果都不理想，原因是多方面的，一方面是目前句法分析的正确率太低，另一方面，目前基于句法的翻译模型不完善，也是重要的原因。不过，现在越来越多的研究者开始意识到这个问题，基于句法的统计翻译模型已经成为了目前的研究热点，相信不久的将来这方面的工作将会有较大的进展。

近期的研究表明，短语的识别和翻译对于跨语言检索也有相当显著的影响。正确的短语翻译可以显著减少翻译的歧义性，缩小检索的范围，提高查询的准确率。因此统计机器翻译在短语处理上的进展对于跨语言检索也有相当正面的影响。不过相对而言，由于跨语言检索对机器翻译有一些特定的要求，简单地将统计机器翻译应用于跨语言检索可能会遇到一些问题，需要将统计机器翻译方法针对跨语言检索的特殊需求进行调整，这方面的研究还有待开展。

4 统计机器翻译与传统机器翻译的结果对比分析

目前在一些公开的机器翻译评测中，统计机器翻译方法已经取得了比较明显的优势。这里，我们通过一些实际的例子来看看这两类系统的翻译效果。

源语言句子：

> 前来出席八国集团同发展中国家领导人对话会议的国家主席胡锦涛16日下午在俄罗斯圣彼得堡同美国总统布什会晤。双方就中美关系和共同关心的重大国际及地区问题深入交换了意见。

Yahoo Babelfish[6]的翻译结果：

> Comes to attend eight countries groups to converse conference state president Hu Jintao with the developing nation leaders on 16th afternoon to meet in the Russian St. Petersburg with American President Bush. Both sides and cared about significant international and the local problem together on the Chinese and American relations have thoroughly exchanged the opinion.

华建在线翻译[7]的结果

> President Hu Jintao coming to attend leader's conference of dialogue of the Group Eight and developing country meets in Sankt Peterburg and US President Bush of Russia on the the afternoon of the 16th. Both sides deeply exchange views on Sino-US relations and great international and regional question of common concern.

Google Language Tools[8]的翻译结果：

> Dialogue with the leaders of developing countries to attend the G-8 meeting Chinese President Hu Jintao on the 16th at St. Petersburg in Russia to meet with U.S. President Bush. Sino-U.S. relations and other issues of common concern on both sides of the major international and regional issues in-depth exchange of views.

我们自己开发的机器翻译系统的翻译结果：

> Attend the G-8 with developing countries leader dialogue meeting of the State President Hu Jintao on the afternoon of 16 in Russia St. Petersburg with US President Bush's meeting. Both sides on Sino-US relations and the common concern of the major international and regional issues, and in-depth exchange of views.

上面的四个系统，除了最后一个系统是作者所在的课题组自行开发的以外，其它三个系统都是在网上可以公开测试的系统，读者不妨一试。这几个系统中，前两个系统都是基于规则的系统，后两个系统都是采用统计方法的系统。其中Yahoo Babelfish采用的就是Systran的系统。

这是一篇真实的新闻稿中的句子，显然，这是一个非常复杂的句子，目前的任何一个机器翻译系统都很难给出很好的译文。实际情况也是如此，上面的四个系统给出的译文，基本上都不是合法的英语句子。每个译文都大致反映了源文句子中一些片段的意思，但完整的意思都不合适。不过，如果我们仔细分析一下，还是能比较出其中的优劣。

我们看看以下几个片段的译文，读者可以比较一下：

(1)前来出席

* Comes to attend
* coming to attend
* attend

* Attend

(2)八国集团

* eight countries groups

* Group Eight

* G-8

* G-8

(3)对话会议

* converse conference .

* conference of dialogue

* Dialogue (with the leaders of developing countries to attend the G-8) meeting

* dialogue meeting

(4)俄罗斯圣彼得堡

* Russian St. Petersburg

* Sankt Peterburg (and US President Bush) of Russia

* St. Petersburg in Russia

* Russia St. Petersburg

(5)中美关系

* the Chinese and American relations

* Sino-US relations

* Sino-U.S. relations

* Sino-US relations

(6)共同关心的重大国际及地区问题

* significant international and the local problem

* great international and regional question of common concern

* issues of common concern (on both sides) of the major international and regional issues

* the common concern of the major international and regional issues

(7)深入交换了意见

* have thoroughly exchanged the opinion

* deeply exchange views

* in-depth exchange of views

* in-depth exchange of views

这里译文的顺序与前面相同。译文中括号括起来的部分是错误插入的词语。

我们可以看到，总体上，后面两个系统的翻译更符合英语通常的习惯。比如说："前来参加"，在英语中，我们通常只翻译成attend就行了，没有必要把"前来"的译文come翻译出来。"八国集团"的译文，虽然前两个系统的译文也不算错，显然G-8是一种更简洁而且是普遍采用的译法。后面几个短语也是这样，虽然前面两个系统的翻译都不一定算错，但后面两个系统的译文显然更符合英文通常的表达方式。

另外有两点情况要说明一下。第一点，统计机器翻译方法的效果与训练语料有很大的关系。上面两个统计机器翻译系统，在训练中都使用了大量的新闻语料，所以对这一类新闻句子的翻译效果比较好。如果被测试的句子与系统训练时的语料不符合，可能效果不会比基于规则的系统好很多。不过，对于统计机器翻译系统而言，要增加训练语料是一件比较简单的事情，而对于基于规则的系统而言，要适应一个新的领域要困难得多。第二点，通常基于规则的系统对于短句子效果会比较好一些，因为短句子通常句法分析成功的正确率比较高，如果句法分析正确，通常得到的英文句子整体结构会比较好，而采用统计方法的系统，由于没有进行句法分析，可能句子结构会差一些，但一些片段的翻译会比较好。

5 结论和展望

综上所述，统计机器翻译的研究近年来取得了非常大的进展，其效果已经远远超过了传统的方法。由于统计机器翻译无需训练即可获得大量细粒度的翻译知识，尤其是对一些短语片段的翻译效果比较好，这特别适合于跨语言检索中查询语句的翻译。

当然，统计机器翻译用于跨语言检索中也会面临一些问题。比如说，平行语料库的获取问题。虽然平行语料库的获得比规则库的编写和调试要容易得多，但要获得大规模的句子对齐语料库，还是一件比较困难的工作，尤其是对于一些比较小的语种来说。另外，语料库的覆盖领域也是很大的问题。网络搜索中查询可能涉及的领域非常广泛，这对语料库的领域覆盖性也提出了很大的挑战。现在有一些研究工作试图从可比语料库，而不是平行语料库中获取翻译知识，这也许是解决这一问题的可行方法之一。可比语料库不要求输入的两种语言的语料完全互为翻译，只要是两个大规模的单语语料库涉及的领域基本相似即可。这大大扩大了语料库的选择范围，但获取双语翻译知识的难度也大了很多。对于在跨语言

检索中应用统计机器翻译来说，语料库的时效性也是一个很大的问题。由于网络搜索中查询语句通常都涉及一些时效性很强的新词或者新短语，而这些新词和新短语通常都还没有出现在语料库中，机器翻译系统通常很难正确地进行翻译。当然，这个问题并不是统计机器翻译所特有的，其它跨语言检索的方法也同样要面临这样的问题。

我们相信，统计机器翻译的成功，必将有力地促进跨语言检索的进展，统计机器翻译研究中形成的一些成功的方法，也一定会对跨语言检索的研究提供有益的经验和教训。

参考文献

[1] W. J. Hutchens. Latest Development in MT Technology: Beginning a New Era in MT zesearch. In: Proceedings of Machine Translation Summit—IV, Kobe, Japan, 1993.

[2] 冯志伟.面向计算机的语言研究.语文与信息,1995(1,2), 1996(3)

[3] 刘群.统计机器翻译综述.中文信息学报,2003,17(4):1—12

[4] NIST 机器翻译评测网站. http://www.nist.gov/speech/tests/mt/

[5] Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical Phrase—Based Translation. HLT/NAACL 2003

[6] Yahoo Babelfish. http://babelfish.yahoo.com/

[7] 华建在线翻译. http://www.hjtrans.com/

[8] Google Language Tools. http://www.google.com/language_tools?hl=zh—CN

作者简介

刘群，男，中国科学院计算技术研究所研究员，博士，研究领域包括：中文自然语言处理、机器翻译和信息提取。通讯地址：中国科学院计算技术研究所　100080

骆卫华，男，中国科学院计算技术研究所助理研究员，在职博士生，主要研究方向是信息检索，信息提取，话题检测与跟踪等。 通讯地址：同上

Application and Advance of Machine Translation Technology in Information Retrieval

Liu Qun, Luo Weihua / Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080

Abstract: This paper introduces the application and advance of machine translation on the background of cross—language information retrieval. At First the usage of machine translation in cross—language information retrieval is introduced. Then a brief history of machine translation, as well as various machine translation methods, especially the recent advance of statistical machine translation is described. The advantages and disadvantages of each method are analyzed. A detailed example is given to show the basic procedure of the state of art of phrase—based statistical machine translation. The translation results of a long news sentence generated by several typical rule—based and statistical machine translation systems is present, and detailed comparison and analysis is given. The last section we have a conclution and prospects the future development of this area.

Keyword: Cross—language information retrieval, Statistical machine translation

(上接 7 页)

参考文献

[1] Miguel E. Ruiz. Cross Language Information Retrieval (CLIR),http://informatics.buffalo.edu/faculty/ruiz/teaching/Seminars/Cross—Language_information_Retrieval.ppt

[2] Douglas W. Oard, Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS—TR—96—19. University of Maryland, Institute for Advanced Computer Studies, http://www.ee.umd.edu/medlab/filter/papers/mlir.ps

[3] Christian Fluhr. Multilingual information retrieval. InRonald A Cole. Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue. Survey of the State of the Art in Human Language Technology, 391—405. Center for Spoken Language Understanding, Oregon Graduate Institute, http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html.

[4] Jian—Yun Nie. Towards a Unified Approach to CLIR and Multilingual IR, Cross—language Information Retrieval: a research Roadmap. [2002—8—14] http://ucdata.berkeley.edu/sigir—2002/

[5] David A. Hull, Gregory Grefenstette. Experiments in multilingual information retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996. http://www.xerox.fr/grenoble/mltt/people/hull/papers/sigir96.ps.

[6] Ballesteros, L., Croft, W.B. Dictionary—based Methods for Cross—Lingual Information Retrieval. In: Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications: 791—801

[7] Chen, Hsin—Hsi. Cross—Language Information Retrieval. In: Proceedings of ROCLING Workshop on ED/MT/IR, Academic Sinica, Taipei. 1997. 4—1~4—27.

[8] Chen, Hsin—His, Lee, Jen—Chang. Identification and Classification of Proper Nouns in Chinese Texts. In: Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996: 222—229.

[9] Ted E. Dunning, Mark W. Davis. Multi—lingual information retrieval. Memoranda in Cognitive and Computer Science MCCS—93—252, New Mexico State University, Computing Research Laboratory, 1993. http://crl.nmsu.edu/ANG/MWD/Book2/mitr.ps.gz.

[10] Chung hsin Lin and Hsinchun Chen. An automatic indexing and neural network approach to concept retrieval and classi_cation of multilingual (Chinese—English) documents. IEEE Transactions on Systems, Man and Cybernetics.1996, 26(1): 75—88. http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html.

[11] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6):391—407. http://superbook.bellcore.com/std/papers/JASIS90.ps.

[12] Berry, M.W., Young, P.G. Using Latent Semantic Indexing for Multilingual Information Retrieval. Computers and Humanities, 29(6):413—429

作者简介

骆卫华，男，中国科学院计算技术研究所助理研究员，在职博士生，主要研究方向是信息检索，信息提取，话题检测与跟踪等。 通讯地址：中国科学院计算技术研究所　100080

An Introduction to Algorithms of Cross-Language Information Retrieval

Luo Weihua / Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080

Abstract: This paper introduces the origin and development of cross—language information retrieval (abbr. CLIR), analysizes key issues of CLIR by discussion of single language information retrieval, and describes separately three main techniques based on lexicon, corpora or modules of machine translation. Finally, we briefly review some new ideas such as a unified method of CLIR and CLIR evaluations.

Keywords: Cross—language information retrieval, Lexicon, Corpora, Machine translation