《当代语言学》第 11卷 2009年第 2期 147－158页，北京

机器翻译研究新进展*

刘群　　中国科学院计算技术研究所

提要　本文介绍近年来国际机器翻译研究领域取得的一些进展，着重介绍统计机器翻译方面取得的进展。具体包括：统计机器翻译的原理和特点、统计机器翻译的发展历程和现状、基于词的统计机器翻译方法、基于短语的统计机器翻译方法、基于句法的统计机器翻译方法等。最后对机器翻译研究今后的发展进行了讨论和展望。

关键词　统计机器翻译　基于词的方法　基于短语的方法　基于句法的方法

1. 概述

机器翻译的发展，经历了一个曲折的过程。按照冯志伟（1994，1996）的说法，到 1980年代，机器翻译研究经历了草创期、萧条期、复苏期、繁荣期等几个阶段。草创期以 1954年在美国乔治敦大学用 IBM计算机进行的首次机器翻译实验为标志，这时的机器翻译方法还比较简单，基本上采用的是单纯的查词典和词频统计等方法，笔者这里称之为朴素的统计方法。而 1964年的 ALPAC报告将全世界的机器翻译热潮打入了冷宫。在 1970年代，随着 Chomsky语言学的兴起和人工智能研究的发展，人们普遍认为要实现机器翻译必须对语言进行理解，在这种背景下，基于规则的机器翻译方法开始发展起来。进入 1980年代以后，机器翻译进入繁荣期，基于规则的机器翻译方法逐步成熟，市场上出现了很多机器翻译系统。但这种繁荣并没有持续下去。1980年代末期到 1990年代，人们发现基于规则的机器翻译系统性能很难进一步提高，面向社会生活中使用的真实语言的时候，机器翻译系统几乎无法给出有用的译文。笔者把这个阶段称为平台期。也正是这个阶段，基于语料库的机器翻译方法开始被提出来并取得了一些进展，这包括基于实例的方法和基于统计的方法。而从 1999年开始到现在，统计机器翻译方法取得了突破性的发展，并且现在仍在迅速发展之中。这个阶段可以称为再度繁荣期。本文着重介绍近年来统计机器翻译的发展状况以及取得的一些新进展。

1.1 统计机器翻译的基本原理和特点

在统计机器翻译中，“翻译”被理解为一个随机事件。也就是说，将一个源语言句子翻译成目标语言句子的过程是完全随机的。更一般地说，可以认为，任意目标语言句子 E都可以是任意源语言句子 F的译文，只是概率不同而已。这样，我们就可以定义一个概率 P(E|F)，用于描述将 F翻译成 E的概率。这个概率被称为翻译模型。翻译模型要满足归一化条

* 本文得到国家高技术研究开发计划（863）项目课题（2006AA010108）和国家自然科学基金项目（60736014和60573188）资助。

件：也就是说，在 F给定的情况下，对于目标语言中所有的句子 E，这个概率之和为 1。这样，机器翻译问题就可以被分解为三个问题：

建模：对 P(E|F)进行定义，给出其数学描述。这是统计机器翻译的核心问题。训练问题和解码问题的解决都是由统计翻译的模型决定的。

训练：利用语料库训练 P(E|F)的参数。

解码：就是翻译。对于给定的句子 F，在译文空间中，搜索概率 P(E|F)最大的句子 E。

我们可以看到，与传统的基于规则或者基于实例的机器翻译方法相比，统计机器翻译理论上具有以下特点：

（1）有严格的数学理论做基础。所有翻译知识，包括词典、规则等等，都以概率的形式呈现，也就是说表现为某种参数。训练过程就是为了得到这些参数，而解码过程就是利用这些参数去搜索最好的译文。在解码过程中，只需要使用这些参数，而不需要再去访问原始的语料库。

（2）不需要人工构造的翻译知识（包括规则和词典），所有语言知识都是从语料库中自动获取。这并不是说，统计机器翻译不需要翻译知识，而是说所有这些翻译知识都是从语料库中自动获取的。目前，统计机器翻译所使用的语料库一般都是双语句子对齐的语料库。语料库的规模通常在几万句对到几百万句对不等。几万句对的语料库通常只能适用于极小的翻译领域，或者仅仅在实验中用来验证某种新的理论或者方法。

（3）翻译的过程被看成是一个最优解的搜索过程。系统从巨大的可能译文空间中寻找最优的译文，搜索的算法采用人工智能中的一些成熟算法。

由于无需人工编写和调试词典及规则，使得统计机器翻译系统在开发和应用上也出现一些明显的特点和优势：（1）机器翻译系统开发的人工成本低、开发周期短；（2）可以迅速迁移到新的语种；（3）可以迅速迁移到新的领域。

1.2 统计机器翻译的发展历程

统计机器翻译的思想，最早是 IBM的研究人员在 1980年代末和 1990年代初提出来的。IBM的研究人员当时使用 IBM最先进的工作站开展了统计机器翻译研究，用短短几年时间、在没有采用任何人工构造的语言知识的情况下，仅仅利用双语语料库，就构造了一个跟 Systran公司历时几十年时间开发出来的法英机器翻译系统相媲美的系统，引起了研究人员的广泛关注。但由于当时计算能力的限制，普通研究人员很难得到 IBM公司那样先进的计算条件，其他研究者也无法重复 IBM公司的工作，以至于这项研究很长时间以来进展非常缓慢。直到 1999年，普通计算机的计算能力已经远远超出了当时 IBM的工作站的水平，在一次约翰霍普金斯大学的夏季研讨班上，一些对统计机器翻译感兴趣的研究人员成功地重复了 IBM当年的工作，并将有关开发工具以开放源代码的形式公开出来，由此引发了统计机器翻译研究的一个新热潮。

下面我们列出统计机器翻译研究中的一些重要历史事件：

1990年代初 IBM首次开展统计机器翻译研究；

1999年 JHU夏季研讨班重复了 IBM的工作并推出了开放源代码的工具；

2001年 IBM提出了机器翻译自动评测方法 BLEU;

2002年 NIST开始举行每年一度的机器翻译评测;

2002年第一个采用统计机器翻译方法的商业公司 Language Weaver成立;

2002年 Och提出统计机器翻译的对数线性模型; (Och and Ney 2002)

2003年 Och提出对数线性模型的最小错误率训练方法; (Och 2003)

2004年 Koehn推出 Pharaoh(法老)标志着基于短语的统计翻译方法趋于成熟; (Koehn 2004)

2005年 Chiang提出层次短语模型并代表 UMD在 NIST评测中取得好成绩; (Chiang 2005)

2005年 Google在 NIST评测中大获全胜, 随后 Google推出基于统计方法的在线翻译工具, 其阿拉伯语-英语的翻译达到了用户完全可接受的水平;

2006年 NIST评测中 USC-ISI的树到串句法模型第一次超过 Google(仅在汉英受限翻译项目中);

2007年 Google宣布推出采用统计机器翻译技术的跨语言搜索网站。

从上面的介绍中可以看出, 在统计机器翻译的发展过程中, 公开的技术评测起到了非常重要的推动作用。图 1给出了国际上最著名的 NIST评测中最近几年的最好成绩:

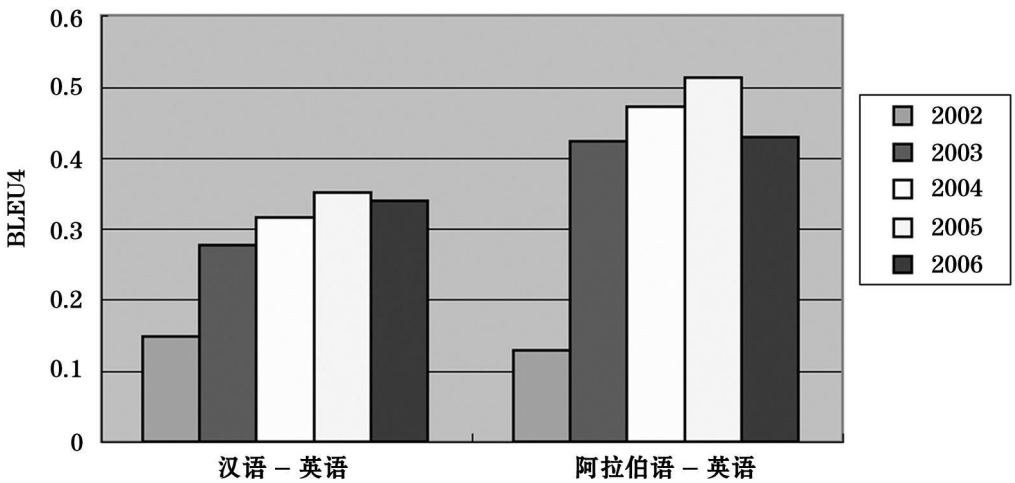

图 1 NIST评测最好成绩

图 1中反映机器翻译质量的一种机器翻译自动评测指标 BLEU。BLEU值本身跟测试集密切相关, 只有相对的比较意义。由于每一年测试的测试题不同, 因此, 每年的结果理论上是不可比的。由于从 2002年到 2005年, 每年的测试题从领域、规模、参考译文数量上看都大致相同, 因此这几年的成绩还是有一定可比性的。而 2006年评测的难度有较大幅度的增加, 主要是测试的领域范围有所扩大, 从单纯的新闻文本扩充到了网络文本和电视访谈记录文本, 规模也从 1000句左右扩大了将近一倍。

从图 1中可以看出, 每年评测的 BLEU值都有所上升, 只是 2006年评测的最高 BLEU值略有下降。考虑到 2006年测试题的难度有较大增加, 可以认为, 这些年统计机器翻译的水

平总体上呈现出了比较明显的上升趋势，这也是统计机器翻译研究吸引人的地方所在。

由于统计机器翻译研究采用了大量的公开的训练数据和测试数据，而且很多研究机构都把他们的研究成果以开放源代码的形式公开出来，这样就大大降低了机器翻译研究的门槛，使得越来越多的研究机构和研究人员投入到机器翻译的研究当中，这也是统计机器翻译这些年发展如此迅速的原因。

1.3 统计机器翻译目前的水平

由于 NIST评测的引导，目前国际上统计机器翻译最主要集中在阿拉伯语到英语和汉语到英语这两个翻译方向上。从目前的研究现状看，阿拉伯语到英语的翻译准确率已经达到了很高的水平，而汉语到英语的翻译准确率相对来说要低一些。下面给出我们利用 Google Translate(http://translate.google.com)翻译的两段即时新闻，以便读者可以对目前统计机器翻译的水平有个大致的了解。这两段新闻一段原文是阿拉伯文，来自阿拉伯半岛电视台，另一段原文是中文，来自新浪网：

(1) مئات الأشخاص شاركوا بمظاهرات احتجاجا على زيارته

Hundreds of people took part in protest demonstrations on his visit

الرئيس الأميركي يؤكد من البحرين اهتمامه بأمن الخليج

American President confirms interest in the security of Bahrain Gulf

بوش وصل المنامة قادما من الكويت بجولة خليجية ستقوده لأبو ظبي والرياض (الأوروبية)

Bush arrived in Manama from Kuwait Gulf tour will be led to Abu Dhabi and Riyadh (European).

شدد الرئيس الأميركي جورج بوش بالمنامة ثاني محطة له في جولته الخليجية على ضرورة تعزيز أمن الخليج

The American President George Bush in Manama second station in the Gulf tour on the need to enhance the security of the Gulf

(2) 中新网 1月 12日电据法新社报道，五角大楼一名官员本周五宣布另外两起美国伊朗军舰对峙事件，称去年 12月时伊朗快艇曾两次在霍尔木兹海峡接近美国军舰，其中一次美国军舰开火示警。

— January 12 according to Agence France-Presse reported, the Pentagon announced that a five-week officials from the two other warships confrontation between the USA and Iran, said last December when the Iranian speedboats twice in the Strait of Hormuz close to the United States Warships, including one American warships fired warning

这名官员称去年 12月 19日，一艘美国两栖战舰在通过霍尔木兹海峡时，因遭伊朗快艇逼近，开火示警。

The official said that on December 19 last year, an American amphibious ships through the Strait of Hormuz, Iran because of the approaching speedboat, fired warning

2. 统计翻译模型的发展

前面我们介绍了，统计机器翻译的一个核心问题就是定义统计翻译模型，也就是一个源语言句子 F翻译成一个目标语言句子 E的概率： P(E|F)。这样翻译问题就变成了一个搜索最大概率的问题：

$$E = \arg\max_{E} P(E \mid F)$$

IBM公司的研究人员（Brown 1990）提出了一种统计机器翻译框架——信源信道模型，将以上翻译模型替换为一个语言模型和一个反向翻译模型的乘积：

$$E = \arg\max_{E} P(E) \times P(F \mid E)$$

这里，P(E)通常被称为“语言模型”。语言模型可以理解成为一种语言的概率化定义。在这种定义下，给定一个单词表，我们认为由单词表中的单词组成的任何有限长度序列都可以是这个语言中的句子，只是概率不同，而这个概率由 P(E)来定义。直观地理解，可以认为语言模型就是刻画了一个句子的流利度，或者说，就是这个句子在多大程度上符合该语言的语法和常用的搭配习惯。

语言模型对统计机器翻译的结果有非常重要的作用。如果仅考虑翻译模型而不考虑语言模型，我们得到的译文很可能是忠实度比较好，即能够将词语翻译正确，但流利度很差，这通常表现为词序比较混乱。而语言模型的引入，使得统计机器翻译结果的可读性大大提高。目前，研究人员所采用的语言模型通常都是 n元语法模型，这是一种比较简单的模型，但效果很好。

2002年，著名的统计机器翻译学者 Och提出了一种新的统计机器翻译框架——对数线性模型（Och and Ney 2002）。该框架比 IBM提出的信源信道模型更具有一般性，信源信道模型可以认为是对数线性模型的一个特例。在对数线性模型框架下，除了翻译模型和语言模型，我们还可以引入任意对机器翻译有贡献的特征函数（比如词典特征、句子长度特征等），而且各个特征函数之间还可以通过训练数据调整加权系数，以达到最好的翻译效果。

在统计机器翻译所使用的所有特征函数中，最重要的仍然是翻译模型和语言模型。由于语言模型理论上一直没有太大进展，所以近年来机器翻译取得的进展实际上主要体现在翻译模型的改进上。本节主要介绍这方面的工作。

统计翻译模型的发展，到目前为止，经历了基于词的模型、基于短语的模型和基于句法的模型三个阶段。目前基于短语的模型是最为成熟的模型，而基于句法的模型也是现在的研究热点。与基于规则的机器翻译方法类似，统计机器翻译模型也可以表示为一个类似金字塔的形式，如图 2所示：

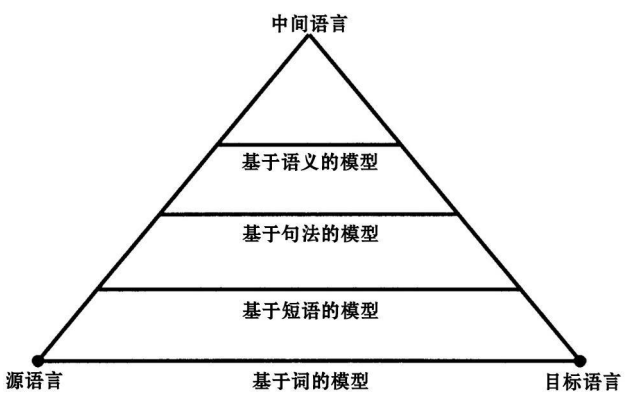

图 2 统计机器翻译模型的金字塔形式示意图

在这个金字塔上，越往塔尖的方向走，对语言的分析也越深入。理论上，对语言的分析越深入，所具有的排歧能力应该越强，译文的质量也应该越高。但实际上，语言的分析本身就是一个很难的问题，分析的深度越深，往往引入的错误也越多，反而会导致翻译质量的下降。因此，如何通过引入更深层的语言分析来提高模型的排歧能力，而又避免分析导致的错

误，就成了统计翻译模型要解决的主要问题。

3. 基于词的统计翻译模型

IBM最早提出的5个翻译模型就是基于词的模型（Brown, et al. 1993），这些模型被称为IBM模型1—5。IBM模型的基本思想是：（1）对于给定的大规模句子对齐的语料库，通过词语共现关系确定词语对齐；（2）一旦得到了大规模语料库上的词语对齐关系，就可以得到一张带概率的翻译词典；（3）通过词语翻译概率和一些简单的词语调序概率，可以计算两个句子互为翻译的概率。

我们看到，这里有一个重要的概念，就是词语对齐。图3给出了一对汉英句子之间词语对齐的一个示意图：

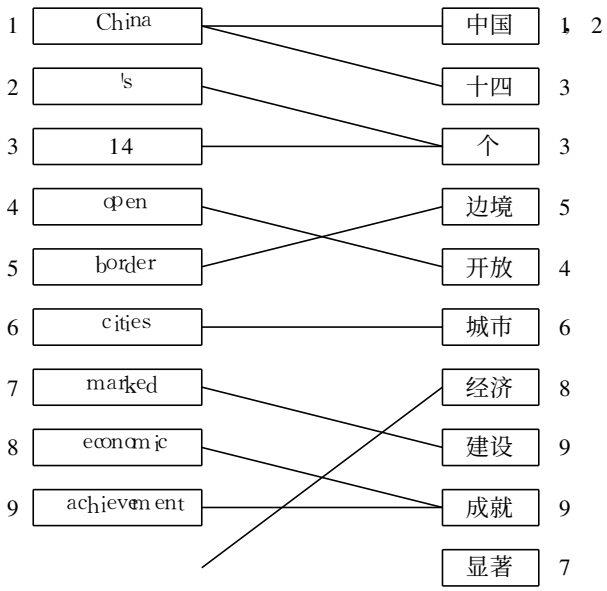

图3 汉英句子词语对齐示意图

IBM模型通过一种很巧妙的方法，可以利用给定的大规模语料库中的词语共现关系，自动计算出句子之间词语对齐的关系，而不需要利用任何外部知识（如词典、规则等），而且可以达到较高的准确率，比单纯使用词典的方法正确率要高得多。其实这种方法的原理也很简单，就是利用词语之间的共现关系。比如说，我们知道以下两个句子对是互为翻译的：

A B ⟷ X Y

A C ⟷ X Z

我们根据直觉，很容易猜想A翻译成X，B翻译成Y，C翻译成Z。只是当我们有成千上万的句子对，每个句子都有几十个词的时候，依靠人的直觉就不够了。IBM模型将人的这种直觉用数学公式定义出来，并给出了具体的实现算法。

通过IBM模型的训练，我们可以利用一个大规模双语语料库得到一部带概率的翻译词典。同时，IBM模型也对词语调序建立了模型，但这种模型是完全不考虑结构的，因此对词语调序的刻画能力很弱，比如说，它可以判断出两个源语言中相邻的词语翻译后依然相邻的

概率较高，如此而已。在基于词的翻译方法中，对词语调序起主要作用的还是语言模型。

在基于词的统计翻译模型下，解码的过程通常可以理解成为一个搜索的过程，或者说，理解成一个不断猜测的过程。这个过程大致如下：

（1）第一步，猜测译文的第一个词，是源文的哪一个词翻译过来的；

（2）第二步，猜测译文的第一个词应该是什么；

（3）第三步，猜测译文的第二个词，是源文的哪一个词翻译过来的；

（4）第四步，猜测译文的第二个词应该是什么；

（5）以此类推，直到所有源文词语都翻译完。

解码的过程中，要反复使用翻译模型和语言模型来计算各种可能的候选译文的概率，同时对低概率的候选译文进行剪枝，以避免搜索的范围过大。

IBM模型可以较好地刻画词语之间的翻译概率，但由于没有采用任何句法结构和上下文信息，它对词语调序能力的刻画是非常弱的。而且由于词语翻译的时候没有考虑上下文词语的搭配，也经常会导致词语翻译的错误。

尽管作为一种基于词的翻译模型，IBM模型的性能已经被新型的翻译模型所超越，但作为一种大规模词语对齐的工具，IBM模型仍然在统计机器翻译研究中被广泛使用，而且几乎是不可或缺的。

4 基于短语的统计翻译模型

很多研究者早就意识到了IBM模型的缺陷：没有采用任何句法结构信息，词序的调整是完全盲目的（完全依靠语言模型来评价）。这就导致其词语调序能力非常弱，对于需要大量调整词序的语言对之间的翻译来说，很难取得很好的性能。

于是很多研究者开始尝试在翻译模型中引入句法知识，也就是说开发基于句法的统计翻译模型（Wu 1997; Yamada and Knight 2001）。但这种早期的尝试性研究大多没有取得好的效果。也有很多研究者采用一种更简单的办法，也就是说，并不采用复杂的句法信息，而是考虑把词语捆绑成短语进行翻译。这里所说的短语并非语言学上的短语，而是任何连续的词串。有时，我们把符合语言学定义的短语称为句法短语，而不符合语言学定义的短语称为非句法短语。

基于短语的统计翻译模型经过很多学者的不懈努力（Wang and Waibel 1998; Och and Ney 2004; Zens et al 2002; Koehn et al 2003），目前已经趋于成熟，其性能已经远远超过了基于词的统计翻译模型，即IBM模型。这种模型是建立在词语对齐的语料库的基础上的，其中词语对齐的工作，仍然要依靠IBM模型来实现。但这种模型对于词语对齐来说，是具有非常高的鲁棒性的，即使词语对齐的效果不太好，依然可以取得很好的性能。

基于短语的翻译模型原理非常简单。也就是在词语对齐的语料库上，我们去寻找并记录所有的互为翻译的双语短语，并在整个语料库上统计这种双语短语的概率。

假设我们已经得到如图4所示的两个词语对齐的片段：

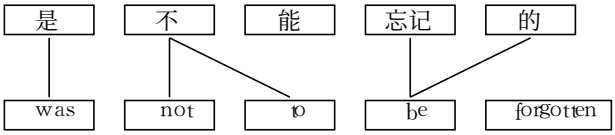

图 4 词语对齐片段示意图

从图 4的小片段中，我们就可以抽取出以下短语：

是	was
是 不 能	was not
是 不 能	was not to
是 不 能 忘记	was not to be forgotten
是 不 能 忘记 的	was not to be forgotten
不 能	not
不 能	not to
不 能 忘记	not to be forgotten
不 能 忘记 的	not to be forgotten
忘记	be forgotten
忘记	to be forgotten
忘记 的	be forgotten
忘记 的	to be forgotten

表 1 短语抽取表

解码 (翻译)的时候，我们只要将被翻译的句子与短语库中的源语言短语进行匹配，找出概率最大的短语组合，并适当调整目标短语的语序即可。

实验表明，非句法短语的作用是巨大的。如果仅仅使用句法短语，系统的性能将会有严重的下降。

读者可能会发现，这种方法几乎就是一种机械的死记硬背式的方法，需要记忆的短语数量是非常庞大的，通常会达到句子数量的几倍到几十倍之多。但对于现有的计算机来说，短语库的规模已经完全不是问题。而且正是这种死记硬背的方法，比一些简单引入句法知识的方法效果好得多。

了解基于实例的机器翻译的读者可能会发现，这种做法很类似于基于实例的翻译方法。其主要区别在于：基于实例的方法通常是翻译的时候到语料库中在线去查找和匹配翻译实例片段，而基于短语的统计翻译模型则是事先将所有双语短语存储起来，翻译时并不需要查找原始的双语语料库，而是只查找双语短语库即可；基于短语的统计翻译模型中所有的双语短语都有概率，系统在所有可能的短语中查找概率最大的组合，而基于实例的方法通常是以相似度作为翻译片段选择的依据。实际上，基于短语的统计翻译系统的性能远远超出了已有的基于实例的机器翻译系统。

5. 基于句法的统计翻译模型

基于短语的统计翻译模型取得了很大的成功，不过其缺陷也是非常明显的。这主要体现在两个方面：（1）模型的泛化能力差。由于采用死记硬背的方法，遇到一些没有在语料库中出现过的搭配就不知道如何翻译了。比如，语料库中只出现了“小王的书、小李的书”，那么在翻译“小张的书”的时候系统就可能不会准确地调整词序了；（2）模型对于长距离语序调整的刻画能力差。短语模型对于短语内部的短距离词语调序有很好的刻画能力，但对于短语之间的语序调整，并没有给出很好的办法。

人们很容易想到要引入句法结构来解决以上问题。不过，研究者们在统计翻译模型中引入句法结构的早期努力大多没有成功。在基于短语的翻译模型成熟以后，研究者们吸收了短语模型的成功经验，开始提出了很多新型的基于句法的统计模型，并取得了初步的成功。这方面的工作我们可以简单分成以下一些类别：

（1）基于形式化句法的统计翻译模型：该类翻译模型建立在形式化句法的基础上。这里所说的形式化句法，指的是从语料库中自动获得的某种句子结构，但并不是通常语言学意义上的句子结构，也不使用任何人类语言学知识，如短语标记（NP、VP）、句子功能关系（主语、谓语）等。这方面的典型工作有（Chiang 2007；Xiong et al. 2006）等。

（2）基于语言学句法的统计翻译模型：该类模型建立在语言学意义上的句法结构基础上，将人类语言学知识包含到模型中。根据所采用的结构树形式的不同，又可以将它分为以下三类：

（a）树到串模型：这一类模型在源语言端引入语言学意义上的句法结构，但在目标语言端并不引入语言学意义上的句法结构。这一类模型的典型工作有（Liu et al. 2006；Quirk et al. 2005）等；

（b）串到树模型：这一类模型在目标语言端引入语言学意义上的句法结构，但在源语言端并不引入语言学意义上的句法结构。这一类模型的典型工作有（Marcu et al. 2006；Galley et al. 2006）等；

（c）树到树模型：这一类模型试图在源语言和目标语言两方面同时引入语言学意义上的句法结构。这一类模型目前还没有比较成功的尝试。

由于篇幅所限，我们在这里仅介绍两个模型。一个是蒋伟提出的基于层次化短语的模型（Chiang 2007），这是一种基于形式化句法的统计翻译模型。另一个是刘洋提出的基于树到串对齐模板的统计翻译模型（Liu et al. 2006），这是一种树到串形式的基于语言学句法的统计翻译模型。

5.1 基于层次短语的统计翻译模型

通过考察基于短语的翻译模型得到的短语翻译概率表，我们会发现，很多短语都是存在嵌套关系的。比如，假设我们的语料库中有以下句子：

（3）澳大利亚 是 与 北 朝鲜 有 外交 关系 的 少数 国家 之 一 。
Australia is one of the few countries which have diplomatic relations with North Korea.

通过这一对句子，我们可以抽取出以下三个短语：

(3') (a) 北朝鲜 ⇔ North Korea
(b) 外交关系 ⇔ diplomatic relations
(c) 与北朝鲜有外交关系 ⇔ have diplomatic relations with North Korea

显然，这三个短语是有嵌套关系的，通过这种嵌套关系，我们可以抽取出一条规则：

(3'') 与 X_1有 X_2⇔ have X_2 with X_1

实际上，这种嵌套关系在语料库中是普遍存在的。根据这种嵌套关系，仅在上面这个句子对中，我们还可以抽取出以下规则：

(3''') (a) X_1之一 ⇔ one of X_1
(b) 是 X_1⇔ is X_1
(c) 是 X_1之一 ⇔ is one of X_1

如果我们用这些规则来取代短语，翻译模型的刻画能力将大大增强，这种模型就是基于层次短语的翻译模型。在这种模型中，没有变量的规则就是短语，因此这种模型与基于短语的模型是完全兼容的，其性能也远远超过了基于短语的模型。

5.2 基于树到串对齐模板（TAT）的统计翻译模型

树到串对齐模板（TAT）的形式如图 5所示：

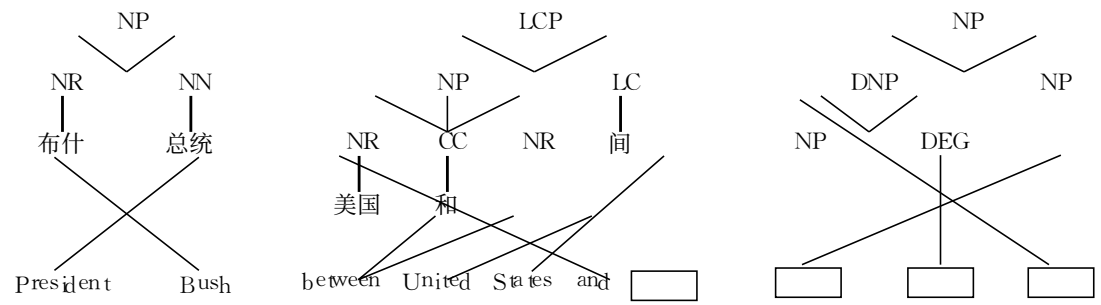

图 5 树到串对齐模板（TAT）示意图

可以看到，在这种模板中，源语言端引入了语言学意义上的句法树，而目标语言端仅仅是词语的序列。因此，采用这种模型进行翻译的时候，要对源语言句子进行句法分析。

有人可能会说，这种模板与传统的基于规则的机器翻译方法中所采用的规则非常相似。确实如此。不过，区别在于，这里的模板是从语料库中全自动地获取的，完全无需人工干预。而且，这里学习得到的模板是带概率的，有完善的统计模型作为理论基础，已经成熟的搜索算法用于寻找最优译文。

由于在自动抽取短语的过程中，我们要考虑所有的词语或者短语（非终结符）都可以表示为变量，因此，这种组合的可能性数量是极为庞大的。根据同样的语料库，我们抽取得到的 TAT模板的数量甚至比双语短语数量还要多很多倍。因此，在实际操作中，我们通常都要通过某种限制来减少模板的数量。

另外，我们可以看到，这种模型非常依赖句法分析的正确率。如果句法分析的正确率不

高，系统的性能也很难提高。而且，由于模型无法利用非句法短语，这对系统的性能也会造成一定的损失。在这些方面，我们通过对现有模型的一些改进，都可以弥补。

基于树到串对齐模板的翻译模型在实践中也取得了很好的效果，性能比基于短语的模型也有了较大幅度的提高。

6 讨论与展望

近几年机器翻译在统计方法的推动下，有了很大的进步，涌现出了很多的新理论和新方法。不仅翻译质量较传统的规则方法有了较大的提高，而且由于可以从大规模语料库中自动获取翻译知识，无需人工撰写规则，大大缩短了机器翻译系统的开发周期，拓展了机器翻译的应用，也降低了机器翻译研究的门槛，这吸引了更多的研究者投入到机器翻译研究中来，使得这个研究领域充满了生机与活力。

到目前为止，统计机器翻译中用到的语言知识还是很有限的。基于词的方法和基于短语的方法几乎没有用到任何语言知识，而是采用了一种词汇化的概率计算方法，所有的语言知识直接通过对词语的概率统计表现出来。目前基于句法的统计翻译方法开始成为研究的热点，特别是在基于语言学句法的翻译模型中，句法知识得到了充分的利用，已经开始超越单纯基于短语的方法或者基于形式化句法的方法。基于语义的方法现在还很少有人用，仅有一些基于词义排歧（WSD）的工作（Chan et al. 2007; Carpuat and Wu 2007），他们的工作证明，词义排歧可以使得现有的机器翻译性能略有提高。

应该看到的是，如果不引入更复杂的语言知识，一些机器翻译问题是不可能真正得到解决的。比如说译文的句法合法性问题、指代问题、篇章问题等等，目前都没有得到很好的解决，这都有待于研究工作者进行更加深入的研究。我们也相信，随着研究的深入，更多的语言知识将能够有效地融入到统计机器翻译之中，使得机器翻译的水平更上一个台阶。

参考文献

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. Proceedings of the Workshop on Speech and Natural Language-ACL, Pp. 146—51.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19, 2, 263—311.

Carpuat, M. and D. K. Wu. 2007. Improving statistical machine translation using word sense disambiguation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007). Pp. 61—72.

Chan, Y. S., H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007). Pp. 33—40.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. Proceedings of ACL 2005. Pp. 263—70.

———. 2007. Hierarchical phrase-based translation. Computational Linguistics 33, 2, 201—28.

Galley, M., J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. 2006. Scalable in-

ference and training of context-rich syntactic translation models. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of Association for Computational Linguistics (ACL 2006). Pp. 961—8.

Koehn, P. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA 2004). Pp. 115—24.

Koehn, P., F. J. Och and D. Marcu. 2003. Statistical phrase-based translation. Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference. Pp. 127—33.

Liu, Y., Q. Liu and S. Lin. 2006. Tree-to-string alignment template for statistical machine translation. Proceedings of COLING/ACL 2006. Pp. 609—16.

Marcu, D., W. Wang, A. Echihabi and K. Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. Proceedings of EMNLP 2006. Pp. 44—52.

Och, F. J. and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. Proceedings of ACL 2002. Pp. 295—302.

———. 2003. Minimum error rate training in statistical machine translation. Proceedings of ACL 2003. Pp. 160—7.

———. 2004. The alignment template approach to statistical machine translation. Computational Linguistics 30.4: 417—49.

Quirk, C., A. Menezes and C. Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. Proceedings of ACL 2005. Pp. 271—9.

Wang, Y.-Y. and A. Waibel. 1998. Modeling with structures in statistical machine translation. Proceedings of COLING/ACL 1998. Pp. 1357—63.

Wu, D. K. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics 23: 377—404.

Xiong, D. Y., Q. Liu and S. X. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. Proceedings of COLING/ACL 2006. Pp. 521—8.

Yamada, K. and K. Knight. 2001. A syntax-based statistical translation model. Proceedings of ACL 2001. Pp. 523—30.

Zens, R., F. J. Och and H. Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, G. Lakemeyer eds., Annual German Conference on AI: KI 2002. Vol. LNAI 2479. London: Springer. Pp. 18—32.

冯志伟，1994，《自然语言机器翻译新论》。北京：语文出版社。

———，1996，《自然语言的计算机处理》。上海：上海外语教育出版社。

作者简介

刘群，男，博士，中国科学院计算技术研究所研究员。研究兴趣：自然语言处理、机器翻译、汉语词法分析与句法分析。电子邮件：liuqun@ict.ac.cn

LIU Qun, mail Ph.D., is a research professor at the Institute of Computing Technology, Chinese Academy of Sciences. His research interest includes natural language processing, machine translation, Chinese lexical analysis and parsing. E-mail: liuqun@ict.ac.cn

作者通讯地址：100080 北京 海淀区中关村科学院南路 6号 中国科学院计算技术研究所智能信息处理重点实验室

E-mail: liuqun@ict.ac.cn

Keywords: Chinese functional words, language knowledge base, attributes of usage

QU Weiguang, TANG Xuri and YU Jingsong, Research on deep processing technologies for large-scale corpora

This paper first examines critically the existing automatic proofreading technologies used in processing Chinese characters. It holds a distinction between shallow tagging and deep tagging. Shallow tagging refers to the use of the existing POS taggers to process texts without human correction of errors. Deep tagging, on the other hand, refers to the method of automatic tagging that improves shallow tagging. The proposed technology has been tested and is found able to detect and correct more than 50,000 errors or inconsistencies in segmentation and POS tagging using the template corpora. The proposed disambiguation model of PFR_SUM (sum of relative frequency ratio of words in context) shows excellent performance in classification, which detects a large amount of errors from template corpora and improves efficiency in corpora proofreading. The model also performs well in solving more than 400 types of common ambiguities when trained on the proofread template corpora and applied to large-scale corpora.

Keywords: shallow tagging, deep tagging, model of rfr_sum, large-scale corpora

LIU Qun, Recent developments in machine translation research

This paper reviews recent developments in machine translation research, especially the research on statistical machine translation. It examines the basic principles and features of statistical machine translation, the history of statistical machine translation, the word-based statistical machine translation approach, the phrase-based approach, and the syntax-based approach. It concludes by exploring the future development of machine translation research.

Keywords: statistical machine translation, word-based approach, phrase-based approach, syntax-based approach

ZHU Weibin, Linguistic models in speech synthesis: The state of the art

This paper examines the state-of-the-art technology in text processing modular used in speech synthesis systems. It includes three computing models: parser, phonetizer and prosodic structure predictor. Problems associated with phonetic error occurrence, prosodic structure prediction and prosodic description in current systems are analyzed. As solutions to these problems, richer linguistic information is proposed to be integrated into the linguistic models and methodologies for achieving the goals are also explored.

Keywords: speech synthesis, linguistic model, speech parser, phonetizer, prosody prediction