统计机器翻译

熊德意 何中军 刘群

摘要：本文对统计机器翻译进行了系统的介绍，综述了 3 种不同类型的统计机器翻译方法：基于词的方法，基于短语的方法和基于句法的方法。对每种方法，简单介绍了其关键思想和有代表性的模型。

关键词：统计机器翻译，基于词的统计机器翻译，基于短语的统计机器翻译，基于句法的统计机器翻译

1 概述

统计机器翻译，又称为数据驱动（data-driven）的机器翻译。其思想其实并不新鲜。早在 1949 年，韦佛（Weaver）发表的以《翻译》为题的备忘录中就提出：“当我阅读一篇用俄语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已。当我在阅读时，我是在进行解码。”这实际上就是基于信源信道思想的统计机器翻译方法的萌芽。早期的机器翻译系统通常都建立在对词类和词序分析的基础之上，分析中经常使用统计方法。只是后来以乔姆斯基（Chomsky）转换生成语法为代表的理性主义方法兴起后，统计机器翻译方法几乎不再被人使用。上世纪 90 年代初期，IBM 的布朗（Brown）等人提出了基于信源信道思想的统计机器翻译模型，并且在实验中获得了初步的成功，引起了研究者广泛的关注和争议。不过由于当时计算能力等多方面限制，真正开展统计机器翻译方法研究的人并不多，统计机器翻译方法是否真正有效还受到人们普遍的怀疑。

但是，进入 21 世纪以来，在学习、生活和工作中，人们日益发现，不同语言之间的交流越来越频繁。无论是口语还是书面形式的交流，无不对机器翻译提出了更加严峻迫切的要求。而另一方面，计算能力也获得了突飞猛进的进步，互联网的发展和普及，以及双语国家、联合国的多语存档，为我们提供了数以千万句的双语平行语料，这些为统计机器翻译方法奠定了必要的基础。于是，越来越多的研究人员开始投入到统计机器翻译的研究中，并取得了成功（在美国国家标准和技术研究所(NIST)信息部语音组主持的机器翻译国际评测中，从 2002 年到 2005 年，统计机器翻译连续四年取得好成绩[1]），统计方法也逐渐成为国际上机器翻译研究的主流方法之一。

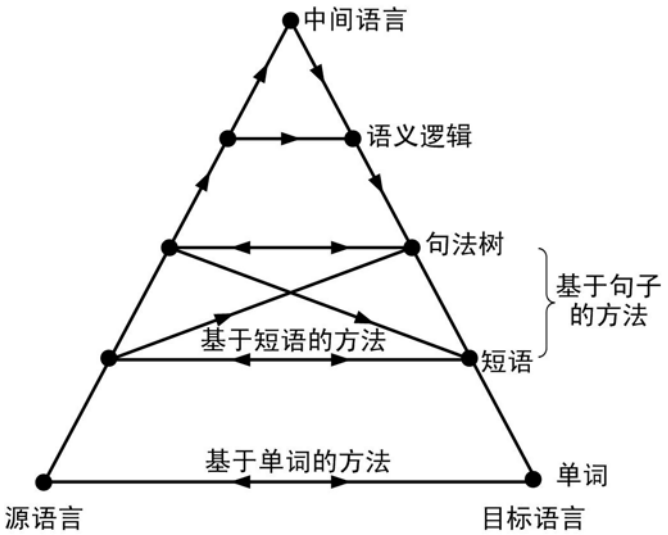


图1. 统计机器翻译金字塔

目前统计机器翻译方法主要分为三类：第一类是基于词的（word-based），以单词作为

翻译的基本单位，不考虑上下文信息和人类语言学知识；第二类是基于短语的(phrase-based)，它将翻译的粒度从单词扩展到短语，能够较好地解决局部上下文依赖问题，提高了翻译的流利度和准确率；第三类是基于句法的（syntax-based），将句法结构信息引入翻译过程，这种方法是当前研究的一个热点。这三类方法可以用图 1 的金字塔来描述。

本文的内容也按照上面提到的 3 种主要方法组织，对每种方法进行简单介绍。由于篇幅所限，我们不可能对每一部分深入展开，但是我们希望通过丰富的实例和图解，以及对翻译方法发展变化的描述，能让读者对统计机器翻译有一个大概的印象。如果能够引起兴趣，那本文的目的就真正达到了。在第二节中我们对基于单词的方法简单介绍，这可以说是统计机器翻译真正意义上的开端引领工作。第三节介绍目前主流的统计机器翻译方法：基于短语的方法。第四节介绍基于句法的方法，这是当前统计机器翻译研究的热点。

2 基于单词的方法

基于词的统计机器翻译，顾名思义，其主要思想是以词作为翻译的基本单位。训练时，它从语料库中统计目标语言单词翻译为源语言单词的概率。翻译时，首先查找每个源语言单词所对应的目标语言单词，然后插入、删除目标语言单词，并调整它们的顺序，最后组合成目标语言句子。这种方法的典型代表就是上世纪 90 年代初期IBM公司的布朗等人提出的基于信源信道模型的方法[2][3]，可以说，它为现代统计机器翻译研究的蓬勃发展奠定了坚实的基础。

信源信道模型将统计机器翻译看成一个信息传输的过程：信道的输入是目标语言 e，在经过信道编码以后，输出源语言 f，机器翻译的任务就是将源语言 f 还原（翻译）为目标语言 e，这一过程通常称之为解码。需要注意的是，这种模型与人们通常的认识有所区别。就模型而言，信道的输入是目标语言，而输出则是源语言，实际上在翻译（解码）时，还是将源语言作为输入，目标语言作为输出。

图2. 信源信道模型示例

根据贝叶斯（Bayes）公式，布朗等人提出了统计机器翻译的基本方程式：

$$\hat{e} = \arg\max_{e} \Pr(e) \Pr(f \mid e)$$

其中，$\Pr(e)$ 是目标语言的语言模型，衡量生成的目标语言的合法程度；$\Pr(f \mid e)$ 是翻译模型，衡量目标语言文本翻译为源语言文本的概率。解码的任务就是根据上式找到概率最大的译文。在此基础上，IBM公司的研究人员提出了 5 个复杂程度层层递进的翻译模型，使

用EM算法[1]从句子对齐的语料库中自动学习单词的翻译概率，然后利用动态规划算法进行解码[3]。

IBM引入的统计方法是通用的，功能也比较强大，在法—英翻译上达到了当时基于句法转换的系统的水平[4]。IBM模型的成功，可以说给整个机器翻译界带来了极大的冲击。它不仅使机器翻译研究者重新思考以前的翻译方法，而且也激发了他们对统计方法用于机器翻译的浓厚兴趣[4][5]。

IBM的工作一直延续到 1995 年，之后由于研究经费的原因而被迫中止[4]。但是由于文献[3]中详细记载了IBM方法，后续的研究者在 1999 年约翰.霍普金斯大学（JHU）夏季研讨班上重新实现了IBM模型，并公开了源代码GIZA[2]。之后奥赫（Och）博士在此基础上发布了增强版GIZA++[3]。这些工作为后来统计机器翻译的发展奠定了坚实的基础。

IBM方法可以说是纯粹的单词到单词自动转录方法，除了计算复杂之外，另外一个很大的缺陷在于它只能学习到两种语言单词之间互为翻译的知识，而对单词的上下文语境却不敏感。这就导致了IBM方法在单词层面（word-level）上由于缺乏上下文语境而不能正确选择译文，尤其是不能正确翻译习惯表达、成语等多个单词结合紧密，却不能通过把逐个单词翻译然后拼凑在一起形成译文的源语言串[6]。

3 基于短语的方法

这种方法的基本思想是以短语作为翻译的基本单位。在翻译过程中，不是孤立地翻译每个词，而是将连续的多个词一起翻译。由于扩大了翻译的粒度，基于短语的方法很容易处理局部上下文依赖问题，能够很好地翻译习语和常用词搭配。一般而言，在基于短语的方法中，短语可以是任意连续的字符串，不作语法上的限制。这样可以方便地从词语对齐的双语语料库中自动抽取双语短语翻译[7]。给定一个源语言句子，基于短语的模型翻译过程如下：

1) 对源语言句子进行短语划分；

2) 根据翻译模型翻译每个短语；

3) 对目标短语进行语序调整。

图 3 是基于短语的翻译过程示例。

图3. 基于短语的方法示例

[1] expectation maximization，期望最大化算法

[2] 参见http://www.clsp.jhu.edu/ws99/projects/mt/

[3] 参见http://www.fjoch.com/GIZA++.html.

3.1 训练

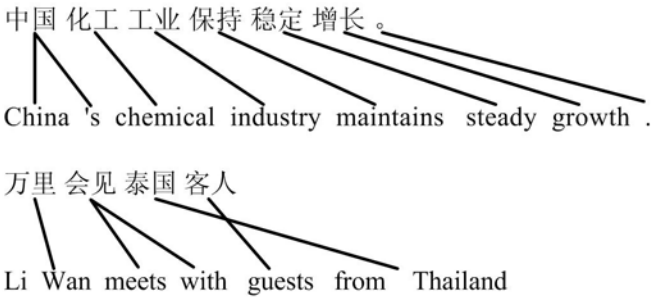
训练的时候，最初输入的是一个双语语料库，即一对一对互为翻译的句子。如样例 1 所示。经过一个词语对齐过程，得到如样例 2 所示形式的数据。从词语对齐的结果中可以知道句子中哪些词是互为翻译的。下面，训练过程的第二个步骤就是进行短语抽取了。所谓短语抽取，就是抽取出语料库中所有互为翻译的连续的词串，而不用管这个词串是否具有真正的含义。以上面第二个句子为例，可以抽取到的短语见样例 3：

样例1

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

样例2



样例3

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

可以看到，“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 operation，而“军事行动”可以翻译成 military action，也可以翻译为 military operation。“大规模杀伤性武器”，绝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”。

3.2 解码

基于短语方法的解码算法比较简单，对于一个源语言句子，它从自动抽取的短语表中查找所有适用的短语翻译，然后选择一个未被翻译的源语言短语进行翻译。翻译信息使用合适的数据结构存储起来，当所有的源语言短语翻译完毕，就完成了解码。许多经典的搜索算法都能用于解码，例如A*算法[23]，动态规划算法[24]等。

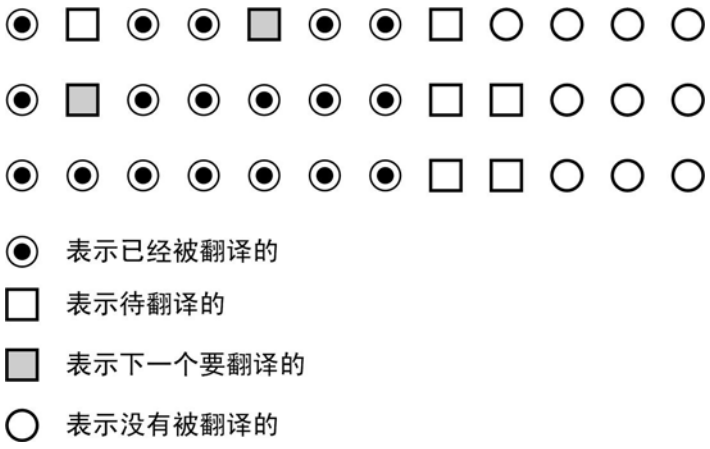


图4. IBM 约束

对于基于短语的方法来说，在解码过程，如何重新排列目标语言短语的顺序（即语序调整），

使译文更加准确、流畅，是一个
很重要的问题。最
简单的方法是不进行短语的语序调整，即译文和原文的短语排列顺序一致，这实际上是完全忽略了语言之间的差异性。但是如果要考虑任意可能的语序调整（即目标短语的全排列），那将是一个NP-难问题[8]。所以实际的做法是引入某种约束条件，减少语序调整的可能性，从而缩小解码时的搜索空间。最常见的两种语序调整约束是IBM约束和ITG约束[9] [12]（见图 4 和图 5）。IBM约束要求每次翻译的源语言短语是最开始没有翻译的K个短语中的一个，这里由于目标语言短语是逐个从左至右组合的，所以目标短语的语序调整问题变成了解码器应该按何种顺序翻译源语言短语的问题。IBM约束规定下一个要翻译的源语言短语不是随意选择的，而是在一个大小为K（可以根据需要定义，K越大，语序调整的可能性就越多，解码复杂度也就越高）的窗口中选择，该窗口涵盖了最开始等待翻译的源语言短语。ITG约束就更好理解了，它要求两个相邻的短语翻译时的顺序只有两种可能性，保序或者逆序，保序是指源语言短语和目标语言短语顺序一致，逆序是指两者的顺序恰好相反。

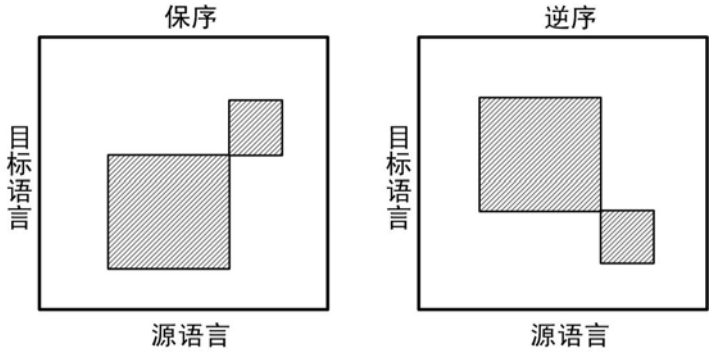


图5. ITG 约束

在这两种约束的基础上，研究者提出了不同的语序调整模型。最开始的语序调整模型与被调整语序的短语是完全不相关的，即独立于短语。后来又发展了和短语相关的模型。

3.2.1 独立于短语的语序调整模型

这类方法在调序的时候不考虑具体的短语，调序带有很大的任意性。主要有以下几种：

1) 固定概率法

这种方法依据具体的两种语言给顺序（Monotone）翻译和非顺序（Non-Monotone）翻译赋予固定的概率，例如顺序翻译的概率是 0.8，非顺序的概率是 0.2。它带有很大的主观性。

2) 移动距离法[7]

这种方法计算短语在源语言和目标语言中移动的距离，赋予惩罚因子，移动距离越大，惩罚也越大。因此，它不鼓励长距离的短语调序。

独立于短语的调序方法比较粗糙，它脱离了具体的短语，调序后短语的顺序主要依靠语言模型来保证，调序效果并不理想。

3.2.2 短语相关的语序调整模型

这类方法在语序调整的时候考虑了具体的短语，从语料库中自动学习短语间的排列顺序，然后应用于翻译过程。由于有了短语信息的指导，调序效果要好于独立于短语的方法。这里简单介绍一下词汇化预测模型[10]。此模型定义了 3 种短语连接方向（Orientation）：左向表示逆序翻译；右向表示顺序翻译；中性方向表示两个短语相对独立。它从语料库中统计相邻短语的连接方向的概率，然后应用于短语排序。该模型的缺陷是，它只能学到相邻短语的排序方式，数据稀疏比较严重。

4 基于句法的模型

基于句法的统计机器翻译最早可以追溯到上世纪 90 年代初，当时同步树粘接语法（Synchronous Tree-adjoining Grammar，简称STAG）[13]和反向转录语法（Inversion Transduction Grammar，简称ITG）[12]相继提出来并用到机器翻译上。从时间上说，与IBM提出基于单词的统计翻译模型的时间很接近，但是基于句法的统计机器翻译研究逐渐得到人们的关注和认可却是在 2000 年之后。即使是在这段时间，许多研究者对基于句法的统计机器翻译仍然持观望态度，这主要是因为基于短语的统计机翻译仍然方兴未艾，最初的基于句法的统计机器翻译系统在性能上与基于短语的统计机器翻译系统相差甚远，再加上 2003 年由奥赫等人组织的约翰.霍普金斯大学夏季研讨班“统计机器翻译的句法方法（Syntax for Statistical Machine Translation）”经过 6 个星期的努力之后，发现引进诸多与句法结构相关的特征并不能显著改善翻译质量[14]。这些使得人们对基于句法的统计机器翻译产生了怀疑[4]。

但是基于短语的统计机器翻译本身存在一些固有缺陷，如短语层次上的全局语序调整，短语非连续性和泛化能力问题，极大地束缚了该方法的进一步发展。这使得人们又不得不求助于句法。因为在理论上或者在人们的自觉上，引入句法结构知识有助于解决这些问题。所以纵观统计机器翻译的发展历程，可以看到，基于句法的统计机器翻译是继基于短语的统计机器翻译之后的一个新趋势。从目前的情况来看，有些基于句法的统计机器翻译系统在性能上已经明显超过了基于短语的系统。如 2005 年美国国家标准和技术研究所机器翻译评测中蒋伟（David Chiang）的 Hiero 系统，2006 年 NIST 机器翻译评测中的 ISI 的系统和中科院计算所的系统，在性能上它们接近甚至超过了最好的短语系统。

将句法知识引入到统计机器翻译系统中，存在多种不同的方法，如在单词对齐模型中引入句法知识[15]，在翻译之前利用句法知识调整源语言语序[16]，在翻译之后利用句法知识做重排序（reranking）[14]等，这里只讨论本质上就是基于句法的翻译模型，并称之为基于句法的统计机器翻译。本文延续蒋伟[17]的分类思想，根据翻译模型所依赖的语法是否包含语言学知识将基于句法的统计机器翻译粗略分为以下 2 类：

1) **基于形式化语法的模型**：该类翻译模型建立在形式化语法的基础上，但并不包含人类语言学知识，如短语标记、词与词之间的依赖关系等；相关的工作有[17]，[11]等。

2) **基于语言学语法的模型**：该类模型建立在语言学语法基础上，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两类：

   a) **基于短语结构树的模型**：该类模型通过短语结构树，将短语的句法标记及标记之间的依赖关系等语言学知识引入到翻译过程中；相关的工作有[18][19]等

[4] 这中间存在一个矛盾，即理论上人们普遍认为句法对机器翻译有很大作用，但是实际系统却并没有证实这一点。这说明或者是句法对机器翻译真地不起作用，或者是人们在利用句法信息时在方法上走了弯路。从后来的工作来看，句法信息如何利用是一个关键问题。

b) **基于依存树的模型**：该类模型通过依存树，将词与词之间的依赖约束关系等语言学知识引入到翻译过程中；相关工作有[20][21][22]等。

下面我们简单介绍基于句法的统计机器翻译的两个代表性工作，一个是蒋伟在[17]中提出的基于形式化语法的层次短语模型；另一个是南加州大学信息科学研究所（ISI/USC）提出的基于语言学语法的串到树模型。

4.1 蒋伟的层次短语模型

该模型在语序调整约束方面比 ITG 模型要弱一些，允许的排序可能性更多一些。其规则形式为：

$$X \rightarrow (\alpha,\ \beta,\ \propto)$$

其中 X 为非终结符，α，β 是非终结符和终结符组成的字符串，$\propto$ 是 α，β 中非终结符的一一对应关系。下面给出几个规则的例子：

$$X \rightarrow (yu\ X_1\ you\ X_2,\ have\ X_2\ with\ X_1)$$

$$X \rightarrow (X_1\ zhiyi,\ \ one\ of\ X_1)$$

为了增强模型的鲁棒性，以及与短语模型的做法保持一致（将句子切分为短语，然后翻译各个短语），蒋伟在他的层次短语模型中引入了以下两条规则：

$$S \rightarrow (S_1\ X_2,\ \ S_1\ X_2)$$

$$S \rightarrow (X_1,\ \ X_1)$$

这两条规则称为“glue”规则，其作用是允许模型对源语言进行短语切分，然后顺序合并它们的译文。显然这两条规则类似于 ITG 中的保序规则，如果只允许这两条规则，再加上短语模型中的短语对，蒋伟的层次短语模型就退化为单调解码的短语模型了。

由于该模型是形式化的句法模型，因此不需要对源语言或目标语言做任何句法分析，就可以从平行语料中抽取这些规则。抽取的方法是：

1) 首先从单词对齐的双语语料中抽取短语

$(f_i^j, e_i^{j'})$，

并将这些短语改造为完全词汇化的基本规则：

$X \rightarrow (f_i^j, e_i^{j'})$，

这里短语的定义和抽取方法与短语模型中的定义和方法是一致的；

2) 然后从已抽取的短语中构建带变量的规则：如果一个短语中包含有其他短语，则将该短语里嵌套的子短语替换为变量 X 。

按这种方法获得的同步规则数量非常大，为了降低复杂度，需要定义一些条件来约束同

步规则的抽取。蒋伟采用的主要约束条件有：

1) 规则的长度在源语言端不能超过L（完全词汇化的基本规则不能超过L_1，带变量的规则不能超过L_2）；
2) 规则最多只能含有 N 个非终结符，即变量 X 的数目不能超过 N；
3) 规则源语言端中不允许存在两个相邻的非终结符；
4) 规则必须含有一对对齐的单词。

那么如何在抽取出来的同步语法上构建概率化的模型呢？借鉴基于短语的统计机器翻译中经常使用的对数线性（log-linear）模型[25]，对于从双语语料中抽取出来的每条规则，其分数可以按下式计算：

$$score(X \rightarrow (\alpha, \beta)) = \prod_i f_i(X \rightarrow (\alpha, \beta))^{\lambda_i}$$

其中 f_i 为定义在规则之上的特征，λ_i 为相应特征的权重。蒋伟定义了如下特征：

1) 两个方向上的短语翻译概率 $p(\alpha \mid \beta), p(\beta \mid \alpha)$；
2) 按照[7]中的方法计算的 IBM 模型 1 的概率，即两个方向上的词汇化概率 $p_{lex}(\alpha \mid \beta), p_{lex}(\beta \mid \alpha)$；
3) 短语个数；

而对于“glue”规则，其分数定义为 $exp(-\lambda_g)$，这可以看作是对保序合并两个相邻变量的一种惩罚。

4.2 ISI 模型系列

南加州大学信息科学研究所（ISI/USC）提出了一系列的串到树模型。它们的基本思想是：目标语言端是有短语结构树的，按照噪声信道模型来解释的话，就是目标语言的树经过有噪声的通道后被异化成源语言的串，解码的任务就是将源语言的串还原成目标语言的树。最初，山田（Yamada）等提出的模型，试图通过一些操作去捕捉噪声通道中的异化过程，并对这些操作赋以不同概率；后来盖雷（Galley）等人将这个异化过程用表达能力更强的规则来描述。规则的一端是带变量的源语言串，另一端是目标语言相应的子树结构。马库（Marcu）等人又在盖雷工作的基础上，加入了更多的特征，并且使得模型可以更好地融合基于短语的统计机器翻译模型中的短语。经过逐步的发展，ISI 提出的串到树模型不断得到完善和改进，以至于他们的系统在 2006 年 NIST 评测的汉-英翻译中名列第一。下面将按照上面所述的 3 个不同阶段来介绍 ISI 的串到树模型。

山田等人的工作

山田等最初提出的模型可以看作是 ISI 串到树模型的一个开端。该模型首先对目标语言进行句法分析，得到目标语言的结构树；结构树通过噪声通道时， 每个节点都要经过一系列所谓的“**通道操作**”，使得最终将所有叶子节点串在一起能得到源语言的串。这些通道操作包括：

1) **插入操作**：选择一个源语言单词，将它插到该节点的左边，或右边，或者干脆不做任何插入。
2) **语序调整操作**：对该节点的所有孩子节点（包括被插入单词形成的节点）进行语序

调整，如果有 N 个孩子节点，则要考虑所有 $N!$ 种可能的语序调整。该操作只对非终结符有效。

3) **翻译操作**：将目标语言结构树的叶子节点翻译为相应的源语言单词。

每个操作都有自己的概率，模型通过EM算法估计这些操作的概率。解码的过程，就是从源语言的句子反向搜索出目标语言的结构树，算法上可以通过CYK[5]来实现。

这些操作可以通过同步上下文无关（SCFG）规则来模拟，如在节点 σ 的左边插入单词 $\boldsymbol{w}$，同步 CFG 规则可以表示为：

$$\sigma \rightarrow (w\sigma_1,\ \sigma_1)$$

因此可以说，山田的模型等价于某种 SCFG 模型。但是我们知道 SCFG 同 CFG 一样，是具有一些缺点的，如它能捕捉的上下文是有限的，由此导致的是 SCFG 只能描述单层树结构，而不是多层。这使得语序调整局限于同一个父节点的不同孩子节点之间，大大地限制了模型的表达能力。另外山田的模型是建立在单词的基础上，这也限制了模型的性能。虽然山田在这两方面都做了一些改进，如引入扁平化操作使得多层之间的语序调整得以实现，允许叶子节点是短语等，但是这些改进终究还是在 SCFG 的框架下，因此一些缺陷还是不能从根本上克服。

继山田之后，盖雷和马库等人对串到树的模型做了一些突破性的工作。盖雷一个主要工作是引入了能够描述多层树结构的规则，大大扩展了模型的表达能力。马库的主要工作是对盖雷的模型继续改进，引入更多特征，并使其能与短语兼容。

盖雷等人的工作

该工作的一个基本思想是把模型建立在能够描述多层树结构的转换规则上。为此，首先对目标语言进行句法分析，得到目标语言的结构树；然后将目标语言的结构树和源语言句子对齐，即目标语言叶子节点对应到源语言的单词上，这个只需要将现有的单词对齐投射到树上就行了；然后从这种串到树的对齐中，按照一定的算法，自动抽取一些规则。这些规则表示了短语结构子树和串之间的对应关系，它们可以简单分为三类：

1) 翻译单词或短语的简单规则，如NP-C(NPB(DT(this) NN(address))) → 这个地址
2) 源语言端带有非终结符的规则，如NP-C(NPB(PRP (my) x0:NN)) → 我的x0
3) 合并规则，即源语言端全部由非终结符组成，如VP(x0:VBZ x1:NP-C) → x0x1

在[26]中盖雷抽取的是最小规则，即和该源语言片段、目标语言结构树以及它们之间的对齐保持一致的最小的规则，它们不能再被拆分为其它规则。而在[18]中，作者不仅抽取了最小规则，而且也抽取了由最小规则组成的复合规则，并且发现这些复合规则对系统的性能提升有很大的帮助。从某种程度上说，可以把最小规则类比为单词，而把复合规则类比为短语。我们都知道基于短语的统计机器翻译要明显优于基于单词的统计机器翻译，这是因为短语捆绑了更多的上下文信息，如局部的单词选择和短语内部的单词顺序等。复合规则相对于最小规则而言，也包含了更多的上下文信息，这是复合规则提升性能的主要原因。除此之外，[18]还详细说明了如何估计规则的概率以及 EM 训练过程。

[5] The Cocke-Younger-Kasami algorithm，一种自底向上的动态规划剖析算法，用于确定能否从上下文无关语法生成串，以及如何生成。

解码的过程类似于单语分析的过程，即对源语言端进行“句法分析”，用转换规则右边去匹配源语言，用左边去生成目标语言的结构树。

马库等人的工作

马库等人在[27]中提出了一个和盖雷等人的工作[18]非常类似的串到树的模型：SPMT[6]模型。虽然该模型表达能力不及盖雷模型强[7]，但马库等人在ISI串到树系列模型上引入了一些新工作，主要有：1）使模型能够兼容非句法短语；2）采用更多的特征函数估计规则的概率。

由于在抽取规则时做出了如下限制：如果一个节点包含在某条规则中，那么该节点所有的姐妹节点也包含在该规则中。这就表明了抽取算法不能抽取那些没有和目标语言结构树对齐的短语。这些短语由于不符合句法，所以称之为非句法短语。非句法短语或者跨越两个不同的句法短语，但是又没有完全覆盖这两个句法短语（比如“布什总统 发表”，“布什总统”是一个完整的名词短语，“发表”则只是动词短语的一部分），或者是一个句法短语的部分孩子节点组成的短语。在基于短语的统计机器翻译系统中，非句法短语得到了大量的应用，事实证明这些短语对系统性能有很大的影响[7]。所以在基于句法的系统中融合基于短语系统中的短语是非常重要的。[27]中一个工作就是使他们的系统能够兼容非句法短语。他们的做法是：对非句法短语，创建一个伪的、非句法的非终结符来覆盖它，由此构建一条转换规则；同时创建另一条相配对的规则，该规则描述了非句法的非终结符如何与其他真正的非终结符组合成句法树。

另外一个工作是估计规则的概率。除了基于句法的特征之外，[27]借用了大量的基于短语模型中的特征函数，如基于 IBM 模型 1 的词汇化概率，语言模型，单词惩罚等。

5 总结

经过近十几年的发展，统计机器翻译有了长足的进步，但是也还有很多难题需要解决：

1) 丰富的语言学知识的引入和使用问题。目前的统计机器翻译系统难以处理复杂多变的语言现象，有的甚至根本不做处理，比如单复数问题、时态问题、句法结构问题等。

2) 大规模数据的处理和使用问题。统计方法离不开大规模数据的支持，在 2006 年的 NIST 评测中，参评系统使用的语料数量达到几百万句对，Google 更是依靠其庞大的机群，使用了 2T 的数据来训练 7 元的语言模型。一般的公司和研究机构都难以处理如此海量的数据，更不要说普通计算机了。

3) 建模问题。随着研究的深入，现在统计机器翻译的模型有不断复杂化的趋势，极大地增加了计算复杂度。理论上来说，模型应该尽量简单，这对于处理大规模的语料库也是很重要的。

4) 搜索问题。搜索算法是与模型相关的。有些模型，如基于句法的模型，本身比较复杂，其搜索算法的复杂度也很高。同时由于受到计算机处理能力的制约，目前的算法不同程度地进行了参数剪枝，这是以损失机器翻译的质量为代价的。

尽管面临众多的技术难题，人们对统计机器翻译仍然充满了热情和信心，就目前的发展

[6] Statistical Machine Translation with Syntactified Target Phrases

[7] 两个模型在目标语言端都能表示多层树结构，但是在源语言端，SPMT模型要求短语必须是连续的，而盖雷等人的模型[18]可以允许非连续短语存在。

来看，统计机器翻译还有巨大的发展潜力和应用前景。以下几个方面值得关注：

1) 统计方法与其它方法的融合。统计的方法关注语言中的共性现象，忽略个性现象，对语言的灵活性把握不够，而其他的方法例如基于规则的方法对语言现象的处理要好很多，而且有很多商用的系统。无论是从理论研究方面还是工程方面，基于统计的方法都应当借鉴和融合其他方法的优点；

2) 重视语言资源的建设。现在我们可以很容易地获得海量的语料库，但是里面含有很多噪声，对模型影响很大。在研究自动处理的方法时，也应当加强对语料库的人工加工和处理，最好能够提供共享使用，这是一件一本万利的事情；

3) 语言知识的引入。基于句法的方法已经迈出了引入语言学知识的第一步。随着研究的不断深入，在统计机器翻译中继续引入各种语言学知识是大势所趋，例如语义知识等，这可能需要设计新的翻译模型；

4) 高效的搜索算法的设计和实现。可以从实现角度考虑，对现有的算法进行优化处理，如设计更好的数据结构、采用并行化处理等；

5) 评测方法的研究。评测为大家提供了一个比较平台，是机器翻译发展的推动力。从近几年 NIST 评测来看，机器翻译的质量不断提高；另外，评测也是机器翻译发展的领航员，评测方法会影响到机器翻译的研究方法。目前很多系统专门针对评测的指标进行改进，有时候偏离了对技术本身的研究。因此，设计合理、实用的机器翻译评测方法也是一个重要的研究方向。

参考文献

[1] David Geer. Statistical Machine Translation Gains Respect. IEEE Computer Society Press, Volume 38, Issue 10, Pages: 18 - 21, 2005.

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics,1990

[3] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol 19, No.2 ,1993

[4] Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit. A Survey of Current Paradigms in Machine Translation. Technique Report: CS-TR-3961, 1998.

[5] John Hutchins. Retrospect and prospect in computer-based translation. In Proceedings of MT Summit VII, 13th-17th September 1999, Kent Ridge Digital Labs, Singapore, 30–34.

[6] Franz Josef Och. 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates. Thesis.

[7] Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)

[8] Kevin Knight. 1999. Decoding complexity in word replacement translation models. Computational Linguistics, Squibs & Discussion, 25(4).

[9] R. Zens and H. Ney: A Comparative Study on Reordering Constraints in Statistical Machine Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 144-151, Sapporo, Japan, July 2003.

[10] C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In *Proc. Of the 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 557–564, Ann Arbor, MI, June.

[11] Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the ACL-COLING*.

[12] D. Wu. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proc. of the 14th International Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1328– 1334, Montreal, August.

[13] Stuart M. Shieber and Yves Schabes. Synchronous tree-adjoining grammars. In Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING), volume 3, pages 1-6, 1990.

[14] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, Dragomir Radev. Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation. 2003.

[15] Dekang Lin and Colin Cherry. 2003. Word Alignment with Cohesion Constraint. In Proceedings of HLT/NAACL 2003. Companion Volume, pp. 49-51, Edmonton, Canada.

[16] Xia, F. and McCord, M. Improving a statistical MT system with automatically learned rewrite patterns. In Proceedings of Coling 2004.

[17] David Chiang. A hierarchical phrase-based model for statistical ma- chine translation. In Proceedings of the 43rd Annual Meeting of the ACL, pages 263-270, 2005.

[18] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Models. In Proc. ACL-COLING, 2006.

[19] Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), pages 609-616, Sydney, Australia, July.

[20] Lin Dekang. 2004. A path-based transfer model for machine translation. COLING 2004.

[21] Chris Quirk, Arul Menezes and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In Proceedings of ACL 2005.

[22] Yuan Ding and Martha Palmer. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars. The 43rd Annual Meeting of the Association of Computational Linguistics, Ann Arbor 2005. (ACL-2005)

[23] Franz Josef Och, Nicola Ueffing, Hermann Ney. An Efficient A* Search Algorithm for Statistical Machine Translation. In: Data-Driven Machine Translation Workshop, pp. 55-62, Toulouse, France, July 2001

[24] R. Zens, F.J. Och, H. Ney. Phrase-Based Statistical Machine Translation. In: M. Jarke, J. Koehler, G. Lakemeyer (Eds.) : *KI - 2002: Advances in artificial intelligence. 25. Annual German Conference on AI, KI 2002*, Vol. LNAI 2479, pp. 18-32, Springer Verlag, September 2002

[25] F.J. Och, H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In Proc. Annual Meeting of the Association for Computational Linguistics , pp. 295-302, Philadelphia, PA, July 2002.

[26] Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. What's in a translation rule? In Proceedings of HLT-NAACL 2004, pages 273-280, 2004.

[27] Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In Proceedings of EMNLP-2006.

作者简介：

熊德意： 中科院计算技术研究所，博士研究生 dyxiong@ict.ac.cn

何中军： 中科院计算技术研究所，博士研究生

刘　群： 中科院计算技术研究所，博士，研究员