统计机器翻译综述[1]

刘群[2]
（北京大学计算语言学研究所 北京 100871）
（中国科学院计算技术研究所 北京 100080）

摘要：本文综述了基于信源信道思想和基于最大熵思想的统计机器翻译方法并介绍了统计机器翻译的评测方法。基于信源信道的方法将翻译概率表示为一个语言模型和一个翻译模型。而基于最大熵的方法则是利用一系列实数值特征函数的线性组合来求解最优的译文。基于最大熵的统计机器翻译方法比基于信源信道的方法更具有一般性，后者可以看做前者的一个特例。

关键词：统计机器翻译 信源信道模型 最大熵方法
中图分类号：TP391

Survey on Statistical Machine Translation

LIU Qun
(Institute of Computational Linguistics, Peking University, Beijing 100871)
(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)
Email: liuqun@ict.ac.cn

Abstract: The paper gives a survey on three approaches of statistical machine translation and the evaluation methods used in SMT. The basic idea of parallel grammar based approach is to build parallel grammars for source and target languages, which conform the same probabilistic distribution. In the source-channel approach, the translation probability is expressed as a language model and a translation model. In the maximum entropy approach, the optimal translation is searched according to a linear combination of a series of real-valued feature functions. The source-channel approach can be regard as a special case of maximum entropy approach.

Keywords: Statistical Machine Translation, Source Channel Model, Maximum Entropy Method

---

[1] 本文工作受国家重点基础研究计划（973）支持，项目编号是G1998030507-4 和G1998030510。
[2] 刘群，男，1966 年生，中国科学院计算技术研究所副研究员，同时在北京大学计算语言学研究所攻读在职博士学位，研究方向是自然语言处理和机器翻译。

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，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。实际上，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 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。1990 年代初期，IBM 的 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。不过，近年来，随着越来越多的研究人员投入到统计机器翻译的研究中并取得了成功，统计方法已逐渐成为国际上机器翻译研究的主流方法之一。

作者根据所查阅的文献，把基于统计的机器翻译方法大体上分为以下三类：第一类是基于平行概率语法的统计机器翻译方法，其基本思想是，用一个双语平行的概率语法模型，同时生成两种语言的句子，在对源语言句子进行理解的同时，就可以得到对应的目标语言句子。这种方法的主要代表有 Alshawi 的 Head Transducer 模型和吴德恺的 ITG 模型，由于这类方法影响较小，而本文篇幅有限，这里不对这类方法进行介绍。第二类是基于信源信道模型的统计机器翻译方法，这种方法是由 IBM 公司的 Peter Brown 等人在 1990 年代初提出的[4,5]，后来很多人都在这种方法的基础上做了很多改进工作，这也是目前最有影响的统计机器翻译方法，一般说的统计机器翻译方法都是指的这一类方法。第三类是德国 Och 等人最近提出基于最大熵的统计机器翻译方法[9]，这种方法是比信源信道模型更一般化的一种模型。

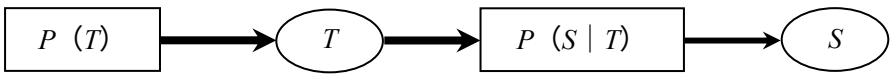
本文将依次介绍后两类统计机器翻译方法，然后介绍一下在统计机器翻译中经常使用的机器翻译自动评测技术，最后给出总结。

2 基于信源信道思想的统计机器翻译方法

2.1 IBM 的统计机器翻译方法

2.1.1 基本原理

基于信源信道模型的统计机器翻译方法的基本思想是，把机器翻译看成是一个信息传输的过程，用一种信源信道模型对机器翻译进行解释。假设一段目标语言文本 T，经过某一噪声信道后变成源语言 S，也就是说，假设源语言文本 S 是由一段目标语言文本 T 经过某种奇怪的编码得到的，那么翻译的目标就是要将 S 还原成 T，这也就是就是一个解码的过程。



根据 Bayes 公式可推导得到：

$$T = \arg\max_{T} P(T)P(S \mid T)$$

这个公式在 Brown 等人的文章[4]中称为**统计机器翻译的基本方程式**（Fundamental Equation of Statistical Machine Translation）。在这个公式中，*P(T)*是目标语言的文本 *T* 出现的概率，称为**语言模型**。*P(S|T)*是由目标语言文本 *T* 翻译成源语言文本 *S* 的概率，称为**翻译模型**。语言模型只与目标语言相关，与源语言无关，反映的是一个句子在目标语言中出现的可能性，实际上就是该句子在句法语义等方面的合理程度；翻译模型与源语言和目标语言都有关系，反映的是两个句子互为翻译的可能性。

也许有人会问，为什么不直接使用 *P(T|S)*，而要使用 *P(T)P(S|T)*这样一个更加复杂的公式来估计译文的概率呢？其原因在于，如果直接使用 *P(T|S)*来选择合适的 *T*，那么得到的 *T* 很可能是不符合译文语法的（ill-formed），而语言模型 *P(T)*就可以保证得到的译文尽可能的符合语法。

这样，机器翻译问题被分解为三个问题：

1. 语言模型 *Pr(t)*的参数估计；
2. 翻译模型 *Pr(s|t)*的参数估计；
3. 搜索问题：寻找最优的译文；

从 1980 年代末开始到 1990 年代中期，IBM 的机器翻译研究小组在统计机器翻译的思想指导下进行了一系列的研究工作[4,5,2]并实现了一个法语到英语统计机器翻译系统。

对于语言模型 *Pr(t)*，他们尝试了采用 n 语法、链语法等语法模型。链语法模型比 n 元语法模型的优点在于可以处理长距离的依赖关系。下面我们着重介绍翻译模型。

2.1.2 IBM 统计翻译模型

对于翻译模型*Pr(f|e)*，IBM公司提出了 5 种复杂程度递增的数学模型[5]，简称为IBM Model 1~5。模型 1 仅考虑词与词互译的概率$t(f_j|e_i)$。模型 2 考虑了单词在翻译过程中位置的变化，引入了参数$Pr(a_j|j,m,l)$，*m*和*l*分别是目标语和源语句子的长度，*j*是目标语单词的位置，*aj*是其对应的源语单词的位置。模型 3 考虑了一个单词翻译成多个单词的情形，引入了产出概率 $\phi(n|ei)$，表示单词*ei*翻译成*n*个目标语单词的概率。模型 4 在对齐时不仅仅考虑词的位置变化，同时考虑了该位置上的单词（基于类的模型，自动将源语言和目标语言单词划分到 50 个类中）。模型 5 是对模型 4 的修正，消除了模型 4 中的缺陷（deficiency），避免对一些不可能出现的对齐给出非零的概率。

在模型 1 和 2 中，首先预测源语言句子长度，假设所有长度都具有相同的可能性。然后，对于源语言句子中的每个位置，猜测其与目标语言单词的对应关系，以及该位置上的源语言单词。在模型 3,4,5 中，首先，对于每个目标语言单词，选择对应的源语言单词个数，然后再确定这些单词，最后，判断这些源语言单词的具体位置。

这些模型的主要区别在于计算源语言单词和目标语言单词之间的连接（Connection）的概率的方式不同。模型 1 最简单，只考虑词与词之间互译的概率，不考虑词的位置信息，也就是说，与词序无关。好在模型 1 的参数估计具有全局最优的特点，也就是说最后总可以收敛于一个与初始值无关的点。模型 2 到 5 都只能收敛到局部最优，但在 IBM 的实验中，每一种模型的参数估计都依次把上一种模型得到的结果作为初始值，于是我们可以看到最后的结果实际上也是与初始值无关的。

下面以模型 3 为例，说明一下从源语言（英语）文本产生目标语言（法语）文本的过程：

1. 对于句子中每一个英语单词 *e*，选择一个产出率 ϕ，其概率为 $n(\phi|e)$；

2. 对于所有单词的产出率求和得到 *m-prime*；
3. 按照下面的方式构造一个新的英语单词串：删除产出率为 *0* 的单词，复制产出率为 *1* 的单词，复制两遍产出率为 *2* 的单词，依此类推；
4. 在这 *m-prime* 个单词的每一个后面，决定是否插入一个空单词 NULL，插入和不插入的概率分别为 *p1* 和 *p0*；
5. 设 ϕ_0为插入的空单词NULL的个数。
6. 设*m*为目前的总单词数：*m-prime*+ ϕ_0；
7. 根据概率表 *t(f|e)*，将每一个单词 *e* 替换为外文单词 *f*；
8. 对于不是由空单词 NULL 产生的每一个外语单词，根据概率表 *d(j|i,l,m)*，赋予一个位置。这里 *j* 是法语单词在法语串中的位置，*i* 是产生当前这个法语单词的对应英语单词在英语句子中的位置，*l* 是英语串的长度，*m* 是法语串的长度；
9. 如果任何一个目标语言位置被多重登录（含有一个以上单词），则返回失败；
10. 给空单词NULL产生的单词赋予一个目标语言位置。这些位置必须是空位置（没有被占用）。任何一个赋值都被认为是等概率的，概率值为 $1/\phi_0$。
11. 最后，读出法语串，其概率为上述每一步概率的乘积。

2.1.3 搜索算法

从上述 IBM Model 3 的介绍中可以看出，对于统计机器翻译而言，搜索算法是一个严重的问题。因为搜索空间一般都是随着源语言句子的大小呈指数增长的，要在多项式时间内找到全局最优解是不可能的。为了在尽可能短的时间内找到一个可接受的译文，必须采用各种启发式搜索策略。

对于搜索问题，IBM 采用一种在语音识别取得广泛成功的搜索算法，称为堆栈搜索（Stack Search），这里不做详细介绍。其他的搜索算法还有柱搜索（Beam Search）、A*搜索等等。

虽然搜索问题很严重，不过 IBM 的实验表明，搜索问题并不是统计机器翻译的瓶颈问题。实际上，统计机器翻译的错误只有两种类型：

1. 模型错误：即根据模型计算出概率最高的译文不是正确译文；
2. 搜索错误：虽然据模型计算出概率最高的译文是正确译文，但搜索算法没有找到这个译文。

根据 IBM 的实验，后一类错误只占所有翻译错误的 5%。

2.1.4 Candide 系统

与传统的基于转换的机器翻译方法相比，我们可以看到 IBM 的统计机器翻译方法中没有使用任何的非终结符（词性、短语类等）。所有的参数训练都是在词的基础上直接进行的。

IBM 的研究者基于上述统计机器翻译的思想，以英法双语对照加拿大议会辩论记录作为双语语料库，开发了一个法英机器翻译系统 Candide [2]。

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		

Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

上表是 ARPA 测试的结果，其中第一行是著名的 Systran 系统的翻译结果，第二行是 Candide 的翻译结果，第三行是 Candide 加人工校对的结果，第四行是纯人工翻译的结果。评价指标有两个：Fluency（流利程度）和 Adequacy（合适程度）。（Transman 是 IBM 研制的一个译后编辑工具。Time Ratio 显示的是用 Candide 加 Transman 人工校对所用的时间和纯手工翻译所用的时间的比例。）

从指标上看，Candide 已经和采用传统方法的商品系统 Systran 不相上下，译文流利程度甚至已经超过了 Systran。

不过，Candide 采用的并不是纯粹的统计模型。实际上，Candide 采用的是也是一种“ 分析－转换－生成 ”的结构。分析阶段使用了形态分析和简单的词序调整，生成阶段也使用了词序调整和形态生成，分析和生成这两个过程都是可逆的。只有在转换阶段使用了完全的统计机器翻译方法。这种做法可以达到三个目的：使隐藏在词语变形之后的英法语对应规则性显示出来；减少了双语的词汇量；减轻了对齐的负担。不过，也正因为这个原因，有人抨击统计机器翻译是“ 石头汤（Stone Soup）”，并认为在这个系统中真正起作用的还是规则方法，因为英法两种语言词序本身相差就不是太大。通过预先的词序调整，两种语言的词序更为接近，这实际上避开了 IBM 统计机器翻译方法的最大问题。

2.1.5 IBM 统计机器翻译方法小结

IBM 提出的统计机器翻译基本方程式具有非常重要的意义。而 IBM 的其他工作只是对这个基本方程式的一种理解。从理论上说，IBM 的模型只考虑了词与词之间的线性关系，没有考虑句子的结构。这在两种语言的语序相差比较大时效果可能会不太好。如果在考虑语言模型和翻译模型时将句法结构或语义结构考虑进来，应该会得到更好的结果。

IBM 提出的统计机器翻译方法在研究者中引起了相当大的兴趣，很多研究者都开展了相关的工作，并取得了一些进展。下面简要介绍其中的一些改进。

2.2 王野翊（Yeyi Wang）在 CMU（卡内基－梅隆大学）的工作

王野翊在他的博士论文[13]中提出了一种对于 IBM 统计翻译模型的一种改进方法。

由于 IBM 的模型完全没有考虑句子的结构信息，这使得人们怀疑 IBM 模型能否在句法结构相差较大的语言对中获得成功。王野翊在他的口语机器翻译实验中也发现，由于德语和英语这两种语言存在的结构差异，导致 IBM 的词对齐模型成为翻译错误的一个重要来源。为此，王野翊提出了一种改进的统计翻译模型，称为基于结构的翻译模型。

这个模型分为两个层次：粗（Rough Alignment）对齐模型和细对齐（Detailed Alignment）模型。首先，源语言和目标语言的短语通过一个粗对齐模型进行对齐，然后短语内的单词再通过一个细对齐模型进行对齐。粗对齐模型类似于 IBM Model 2，席对齐模型类似于 IBM Model ４。

为了在粗对齐阶段实现双语短语的对齐，王野翊引入了一种双语的文法推导算法。在训

练语料库上，通过基于互信息的双语词语聚类和短语归并反复迭代，得到一组基于词语聚类的短语规则。再用这组规则进行句子的短语分析。

王野翊的实验表明，结构的引入不仅使统计机器翻译的正确率有所提高（错误率降低了11%），同时还提高了整个系统的效率，也缓解了由于口语数据的严重缺乏导致的数据稀疏问题。

2.3 约翰霍普金斯大学（JHU）的统计机器翻译夏季研讨班

IBM 提出统计机器翻译方法引起了研究者广泛的兴趣。不过，由于其他人无法得到 IBM 的源代码，而要进行统计机器翻译的研究，首先需要重复 IBM 的统计机器翻译试验，然后才谈得上对它进行改进。这将面临着编码方面巨大的工作量。于是，在 1999 年夏天，很多相关的研究者会聚在约翰霍普金斯大学（JHU）的夏季研讨班上，大家共同合作，重复了 IBM 的统计机器翻译试验，并开发了一个源代码公开的统计机器翻译工具包——Egypt。在这以后，这些研究者回到各自的研究机构，继续开展相关的研究工作，并提出了各种改进的模型，使得统计机器翻译的研究又出现了一个新的高潮。

在约翰霍普金斯大学的 1999 年统计机器翻译夏季研讨班上，研究者们构造了一个基本的统计机器翻译工具集 Egypt，并将该工具集在感兴趣的研究者中间自由散发。在研讨班上，他们使用这个工具集作为试验的平台进行了一系列的实验[1]。

研讨班开始时预期达到的目标如下：

1. 构造一个统计机器翻译工具并使它对于研究者来说是可用的。这个工具集应该包含语料库准备软件、双语文本训练软件和进行实际翻译的实时解码软件。
2. 在研讨班上用这个工具集构造一个捷克语—英语的机器翻译系统；
3. 进行基准评价。这个评价应该包含客观评价（统计模型困惑度）和主观评价（质量的人工判断），并试图使二者互相联系。我们还要产生一个学习曲线，用于显示系统性能如何随着双语语料的数量发生变化。
4. 通过使用形态和句法转录机改进系统性能；
5. 在研讨班最后，在一天之内构造一个新语对的翻译系统。

研讨班最后完全达到了上述目标。除此之外，研讨班还完成了以下实验：提高双语训练的速度，使用双语词典，使用同源词。并构造了一些工具来支持以上实验，包括一个复杂的图形界面用于浏览词对词对齐的结果，一些语料库的准备和分析工具，和一个人工判断的评价界面。

EGYPT 工具包包含以下几个模块：

1. GIZA：这个模块用于从双语语料库中抽取统计知识（参数训练）。
2. Decoder：解码器，用于执行具体的翻译过程（在信源信道模型中，“ 解码 ”就是“ 翻译 ”）。
3. Cairo：整个翻译系统的可视化界面，用于管理所有的参数、查看双语语料库对齐的过程和翻译模型的解码过程。
4. Whittle：语料库预处理工具。

Egypt 是个免费的工具包，其源代码可以在网上自由下载。这为相关的研究工作提供了一个很好的研究基础。

2.4 Yamada 和 Knight 的工作——基于句法结构的统计翻译模型

南加州大学信息科学研究所(ISI/USC)的 Kevin Knight 是统计机器翻译的主要倡导者之一，在统计机器翻译方面做了一系列的研究和推广工作，他也是 JHU 的统计机器翻译夏季讨论班的主要组织者之一[6]。

Yamada, Knight 等人在 IBM 的统计翻译模型的基础上，提出了一种基于句法结构的统计翻译模型[14]。其主要的思想是：

1. IBM 的信源信道模型中，噪声信道的输入和输出都是句子，而在基于句法结构的统计翻译模型中，噪声信道的输入是一棵句法树，输出是一个句子；
2. 在翻译过程中，对源语言句法树进行以下变换：
   a) 对句法树进行扁平化处理（将相同中心词的多层结点压缩到一层）；
   b) 对于源语言句法树上面的每一个结点的子节点进行随机地重新排列（N 个子节点就有 N!种排列方式），每一种排列方式都有一个概率；
   c) 对于句法树任何一个位置随机地插入任何一个新的目标语言单词，每一个位置、每一个被插入的单词都有不同的概率；
   d) 对于句法树上每一个叶节点上的源语言单词翻译成目标语言单词，每一个不同的译文词选择都有不同的概率；
   e) 输出句子，其概率为上述概率的乘积。

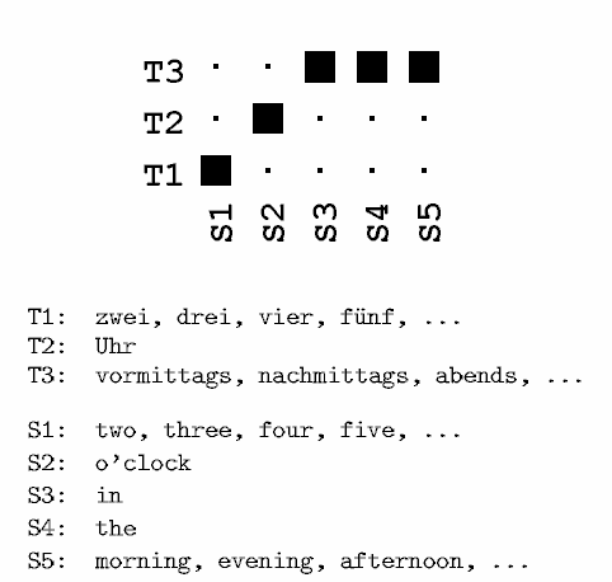
从现有的文章中看，他们的实验采用了一个从英日词典中抽取的例句语料库，一共只有 2121 个句子，平均句长不到 10 个词。虽然其结果比 IBM Model 5 更好，不过由于他们的实验规模还比较小，严格来说并不具有足够的说服力。

2.5 Och 等人的工作

德国 RWTH Aachen – University of Technology 等人在统计机器翻译领域也开展很多的工作。

在德国主持开发的著名的语音机器翻译系统 Verbmobil 中，Och 所在的研究组承担了其中统计机器翻译模块[7]。与 IBM 的模型相比，他们主要做了以下改进：

1. 为了解决数据稀疏问题，他们采用了基于类的模型，利用一种自动的双语词聚类技术，将两种语言的每一个词都对应到一个类中，总共使用了 400 个类；
2. 在语言模型上，采用了基于类的五元语法模型，采用回退（Back-off）平滑算法；
3. 在翻译模型上，采用了一种称为对齐模板（Alignment Template）的方法，实现了两种层次的对齐：短语层次的对齐和词语层次的对齐。对齐模板也采用基于类的对齐矩阵的形式表示，如下图所示：



对齐模板的获取是自动进行的，在对训练语料进行词语对齐以后，所有可能的对齐模板都被保存下来，并根据其在语料库中出现的频率赋予不同的概率。对于一个新句子进行短语匹配的过程类似于一个汉语词语切分的过程，采用一个动态规划算法，寻找概率最大的路径。

4. 为了搜索的方便起见，他们对于 IBM 提出的统计机器翻译基本方程式进行了修改，用一个反向的翻译模型取代了正常的翻译模型，如下所示：

$$S = \max_{S} P(S)P(S \mid T)$$

通过实验他们发现，这种改变并没有降低总体的翻译正确率。

3 基于最大熵思想的统计机器翻译方法

正如上一节所述，Och 等人在进行统计机器翻译实验时发现，把 IBM 统计机器翻译基本方程式中的翻译模型换成反向的翻译模型，总体的翻译正确率并没有降低，这用信源信道理论是无法解释的。于是，他们借鉴了[10,11]中统计自然语言理解的一种思路，提出了基于最大熵的统计机器翻译方法[9]。这是一个比基于信源信道的统计机器翻译方法更为一般化的一种方法，基于信源信道的方法可以看做是基于最大熵的方法的一个特例。

基于最大熵的方法与基于信源信道的方法不同，没有语言模型和翻译模型的划分（虽然也可以将它们作为特征），因而是一种直接翻译模型。

最大熵，又称最大熵原理，或者最大熵方法，是一种通用的统计建模的方法。我们这里简单介绍一下最大熵方法的基本思想[3]。

对于一个随机事件，假设我们已经有了一组样例，我们希望建立一个统计模型，来模拟这个随机事件的分布。

为此，我们就需要选择一组特征，使得我们得到的这个统计模型在这一组特征上，与样例中的分布完全一致，同时又保证这个模型尽可能的“均匀”（也就是使模型的熵值达到最大），以确保除了这一组特征之外，这个模型没有其他的任何偏好。依据这个原则的统计建模方法就是最大熵方法。

假设e、f是机器翻译的目标语言和源语言句子，$h_1(e,f), \ldots, h_M(e,f)$分别是$e$、$f$上的$M$个特征，$\lambda_1, \ldots, \lambda_M$是与这些特征分别对应的$M$个参数（权值），那么直接翻译概率可以用以下公式

模拟（推导略）：

$$\Pr(e|f) \approx p_{\lambda_1 \ldots \lambda_M}(e|f)$$
$$= \exp[\sum_{m=1}^{M} \lambda_m h_m(e,f)] \Big/ \sum_{e'} \exp[\sum_{m=1}^{M} \lambda_m h_m(e',f)]$$

而对于给定的 f，其最佳译文 e 可以用以下公式表示（推导略）：

$$\hat{e} = \arg\max_e \{\Pr(e|f)\}$$
$$= \arg\max_e \{\sum_{m=1}^{M} \lambda_m h_m(e,f)\}$$

可以看到，如果我们将两个特征分别取为log $p(e)$和log $p(f|e)$，并取λ_1＝λ_2＝1，那么这个模型就等价于信源信道模型。

在最大熵方法中最常用的做法是采用二值特征，可以用一种 IIS 算法进行参数训练。而在基于最大熵的统计机器翻译中，由于采用的特征是一种实数值特征，模型的参数不能使用通常 IIS 算法进行训练。为此[Och, 2002]提出了采用了一种区别性学习方法（Discriminative Training），其训练的优化准则为：

$$\hat{\lambda}_1^M = \arg\max_{\lambda_1^M} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(\mathbf{e}_s|\mathbf{f}_s) \right\}$$

这个判定准则是凸的，并且存在全局最优。

Och 介绍了他们在基于最大熵的统计机器翻译方法上的一系列实验[9]：

1． 首先将信源信道模型中的翻译模型换成反向的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降；
2． 调制参数λ_1和λ_2，系统性能有了较大提高；
3． 再依次引入其他一些特征，系统性能又有了更大的提高。

他们引入的其他特征包括：

1． 句子长度特征：对于产生的每一个目标语言单词进行惩罚；
2． 附件的语言模型特征：一个基于类的语言模型特征；
3． 词典特征：计算给定的输入输出句子中有多少词典中存在的共现词对。

可以看到，采用基于最大熵的统计机器翻译方法，确实比简单地采用信源信道模型可以较大地提高系统的性能。

基于最大熵的统计机器翻译方法为统计机器翻译的研究提供了一个更加广阔的视野，这篇论文[9]获得了 ACL2002 的最佳论文奖。

4 统计机器翻译的评测方法

在自然语言处理中，评测的重要性越来越得到人们普遍的重视。在这方面，语音识别研究的进展给了人们很好的启示[8]。现在，在自然语言处理的相关领域，已经出现了一系列以评测带动的学术研讨会，对有关的学术研究都起到了极大的促进作用。

机器翻译评测方法的研究，已经成为机器翻译研究中的一个热点问题。在 MT Summit 2001 上，就有一个专门的 Workshop 讨论机器翻译的评测问题。

机器翻译的评测，主要有人工评测和自动评测两种方法。人工评测的优点是准确率高。缺点是人力成本和时间成本都太高。自动评测的优点是成本低，速度快，可以反复使用。缺点是准确率较低。目前机器翻译评测研究的重点主要在于如何提高自动评测的准确率。

除了公开的机器翻译评测之外，在日常的系统开发工作中，机器翻译评测也是非常重要的。只有通过频繁的评测，才能从各种方法中找到最有效的办法来提高机器翻译的效果。而这种频繁的评测只有自动评测方法才能胜任。

4.1 几个简单的机器翻译自动评价指标

机器翻译自动评价中，经常用到以下几个简单的评价指标：

1. 句子错误率：与参考译文不完全相同的句子的比例。显然，这个指标过于严格。
2. 单词错误率：从候选译文到参考译文，所需要进行的插入、删除、替换操作的次数（每次对一个单词进行操作），除以参考译文中单词的个数；其中，这三种操作的次数之和又称为编辑距离（edit distance）。这个指标是从语音识别中借鉴过来的。由于语音识别的结果语序是不可变的，而机器翻译的结果语序是可变的，显然这个指标存在一定的缺陷。
3. 与位置无关的单词错误率：在单词错误率的计算中，不考虑插入、删除、替换操作的顺序。也就是说，候选译文与参考译文相比，多出或不够的词进行删除和插入操作，其余不同的词进行替换操作。这个指标与单词错误率相比，允许语序的变化，不过又过于灵活。

可以看到，这几个指标都过于简单，对于机器翻译评测来说并不是非常合适。

4.2 基于测试点的机器翻译自动评测方法

1990 年代初俞士汶等[15]提出了一种基于测试点的机器翻译自动评测方法，采用一种类似标准化考试的办法，对机器翻译的各个主要指标设计一定数量的试题进行测试，以达到对机器翻译性能的总体评价。这是一种很巧妙的方法，其优点在于不仅可以实现真正的全自动评测，而且可以按照人的意图实现分项评测，如单独测试系统的词汇能力或者语法能力。这也是世界上较早提出的机器翻译自动评测方案之一。这种方法的主要缺点是试题的编写需要非常专业的人员，成本较高，题库的扩充比较困难。

4.3 IBM 的 BLEU 评价方法

IBM 公司在其一份技术报告[12]中提出了一种基于 n 元语法的机器翻译自动评测方法，其基本思想是，将机器翻译产生的候选译文与人类翻译者提供的多个参考译文相比较，越接近则候选译文的正确率越高。BLEU 是 BiLingual Evaluation Understudy 的缩写，其设想是作为人类专家评测的一个“替身”。

所谓 n 元语法的精确率，就是候选译文中 n 词接续组在参考译文中出现的比例。对于候选译文中某个 n 词接续组出现的次数，如果比参考译文中出现的最大次数还多，要把多出的次数“剪掉”（不作为正确的匹配）。

很容易发现，这种做法只考虑到了“精确率”，而没有考虑到“召回率”。为了避免“召回率”过低的问题，BLEU 的评价标准又对比参考译文更短的句子设计了“惩罚因子”。

在 BLEU 中，n 的实际取值是 1~4。总的评价指标是一元语法到四元语法的几何平均。

另外，对于整个语料库而言，BLUE 的计算是基于词语进行的，而不是基于句子的。也就是说，对于长度不同的句子，要以句子的长度进行加权平均。

因此，BLEU 的总体评价公式如下：

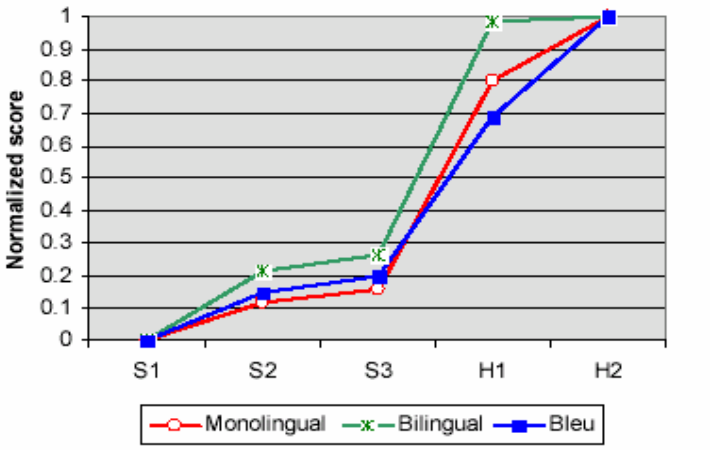
$$\text{BLEU}= \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

其中，p_n是出现在参考译文中的n元词语接续组占候选译文中n元词语接续组总数的比例，$w_n = 1/N$，N为最大的n元语法阶数（实际取 4）。BP为长度过短的惩罚因子，按以下方式计算：

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}.$$

其中 c 为候选译文中单词的个数，r 为参考译文中与 c 最接近的译文单词个数。

根据 IBM 的实验，BLEU 可以相当好地模拟了人类专家对机器翻译的评测结果。参考下图中的曲线。



其中 S1、S2、S3 分别是三个不同的机器翻译系统提供的译文，H1 和 H2 是两个人类翻译者提供的译文。蓝线是 BLEU 系统评测的结果，红线是只懂目标语言的人类专家提供的评测结果，绿线是同时懂源语言和目标语言的人类专家提供的评测结果。可以看到，这三条曲线拟合得相当不错。特别是与只懂目标语言的人类专家相比，在翻译质量不是特别好的时候（恰好现有机器翻译系统的质量都不是太好），曲线的拟合程度更高。

5 总结

前面我们介绍了三种类型的统计机器翻译方法，这里我们把这三种方法做一个简单的比较。

基于平行语法的统计机器翻译方法，虽然已经有了几种不同的模型，不过在实践中并不很成功。我们认为，这种方法虽然理论上比较完善，不过并不符合自然语言的实际情况。因为自然语言之间的差异是非常大的。而在这种方法中，不仅要为两种不同的自然语言的语法之间建立起一一对应的关系，并且还要服从相同的概率分布，这实际上是很不现实的，很难反映自然语言的真实分布规律。

基于信源信道思想的统计机器翻译方法的提出，是统计机器翻译的一个突破。在这种方法中，翻译过程被模型化为一个翻译模型和一个语言模型。通过翻译模型和语言模型的共同作用，使得获得的译文不仅可以忠实的反映源文要表达的意思，而且可以尽可能的做到译文的流畅，取得了较好的效果。不过，就目前的研究而言，翻译模型和语言模型都还显得比较

粗糙，不能反映语言的深层结构信息。

基于最大熵的统计机器翻译方法的提出，大大开阔了统计机器翻译方法的思路。使得我们在机器翻译中不仅可以引入翻译模型和语言模型之外的各种特征，而且可以通过最大熵方法训练出一组参数，找到这些特征之间的一种最优组合形式。我们认为，这种方法的前景是非常广阔的。

IBM 提出的统计机器翻译方法，不仅仅是对机器翻译，而是对于整个的自然语言处理，都产生了长远而深刻的影响。由于各方面的原因，虽然 IBM 的统计机器翻译实验在初期取得了很大的成功（在 ARPA 测试中获得了超过 Systran 的结果），但 IBM 的统计机器翻译工作并没有坚持下来，人们对统计机器翻译的怀疑也始终挥之不去。虽然统计方法在自然语言的很多其他领域都取得了成功，不过在机器翻译领域，统计方法的有效性并没有得到普遍的承认。

应该说，IBM 的统计机器翻译工作是有一定超前性的。IBM 在 1980 年代末到 1990 年代初就开始进行统计机器翻译工作，在 IBM 公布其实验结果之后的很多年，都没有人能够重复类似的结果。随着时间的推移，到了 1990 年代末，相关的研究工作开始得到重视，王野翊、Malamed、Knight、Och 等人都重复了 IBM 的工作并进行了一定的改进。统计机器翻译方法又出现了一个小的高潮。在 JHU 的夏季研讨班的总结报告[1]上有一段话耐人寻味：

> 当解码器的原形系统在研讨班上完成时，我们很高兴地惊异于其速度和性能。在 1990 年代早期在 IBM 公司举行的 DARPA 机器翻译评价时，我们曾经预计只有很短（10 个词左右）的句子可以用统计方法进行解码，即使那样，每个句子的解码时间也可能是几个小时。在早期 IBM 的工作过去将近 10 年后，摩尔定律、更好的编译器以及更加充足的内存和硬盘空间帮助我们构造了一个能够在几秒钟之内对 25 个单词的句子进行解码的系统。为了确保成功，我们在搜索中使用了相当严格的阈值和约束，如下所述。但是，解码器相当有效这个事实为这个方向未来的工作预示了很好的前景，并肯定了 IBM 的工作的初衷，即强调概率模型比效率更重要。

在 2002 年的 ACL 会议上，Och 等人关于统计机器翻译的论文[9]获得了大会最佳论文奖。在 2002 年 NIST 举办的机器翻译评测中，Och 所在的德国 RWTH Aachen – University of Technology 提交的系统获得了最好的成绩，统计机器翻译方法的优势得到了明显的体现。

当然，现在的统计机器翻译方法并不排斥规则方法。从工程实践的角度看，这二者本来就不应该是互相矛盾的。语言学家的理性的思辨帮助我们更加深入的了解语言的本质，并帮助我们建立合理的模型，收集和整理有关的语言数据；而对已有的语言事实和经验数据的有效的统计分析和合理运用，可以使得我们在现有的条件下达到最好的结果。

人类语言知识和统计方法的有效结合，是自然语言处理取得成功的必要条件，这一点应该没有人会再怀疑。不过，在具体的做法上，作者认为有一些基本的原则应该注意把握：

1. 重视大规模语言资源的建设，特别要做到资源的开放与共享；
2. 从研究的角度看，尽量不要去做封闭的、局限于某一具体系统的大规模语言资源（公司行为除外）；
3. 尽可能利用公开的语言资源，其他问题，尽量用统计方法解决；
4. 评测是推动自然语言处理技术发展的有效方法；
5. 评测应该做到公开，即评测规范、评测软件、评测的训练语料都应该公开；
6. 评测应该尽量做到自动进行，减少人的因素。对于评测的组织者而言，这有利于评测的公平性；对于开发者而言，快速、高效、低成本的评测可以有助于迅速选择合适的算法、模型和参数，有效地改善系统的性能。

希望本文对关心和从事机器翻译研究工作的人们有所帮助。

参考文献

[1] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith and David Yarowsky. Statistical Machine Translation: Final Report [R], Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD.

[2] Berger, A., P. Brown, S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz, L Ures, The Candide System for Machine Translation [A], Proceedings of the DARPA Workshop on Human Language Technology (HLT), 1994

[3] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 22(1):39-72,March 1996.

[4] 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 [J], Computational Linguistics,1990

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

[6] Kevin Knight, A Statistical Machine Translation Tutorial Workbook [Z]. unpublished, prepared in connection with the JHU summer workshop, August 1999. (available at http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf).

[7] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation [A]. In Proc. of the Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20-28, University of Maryland, College Park, MD, June 1999.

[8] Franz Josef Och, Hermann Ney. What Can Machine Translation Learn from Speech Recognition? [A] In: proceedings of MT 2001 Workshop: Towards a Road Map for MT, pp. 26-31, Santiago de Compostela, Spain, September 2001.

[9] Franz Josef Och, Hermann Ney, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation [A], ACL2002

[10] K. A. Papineni, S. Roukos, and R. T. Ward. Feature-based language understanding [A]. In European Conf. on Speech Communication and Technology, pages 1435-1438, Rhodes, Greece, September, 1997

[11] K. A. Papineni, S. Roukos, and R. T. Ward. Maximum likelihood and discriminative training of direct translation models [A]. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, pages 189-192, Seattle, WA, May, 1998

[12] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation [R], IBM Research, RC22176 (W0109-022) September 17, 2001

[13] Ye-Yi Wang, Grammar Inference and Statistical Machine Translation [D], Ph.D Thesis, Carnegie Mellon University, 1998

[14] K. Yamada and K. Knight, A Syntax-Based Statistical Translation Model [A], in Proc. of the Conference of the Association for Computational Linguistics (ACL), 2001

[15] 俞士汶等, 机器翻译译文质量自动评估系统 [A], 中国中文信息学会 1991 年会论文集, PP314~319