语言文字应用（2002 年第 1 期）pp77-84

统计语言模型能做什么？

黄昌宁
（微软亚洲研究院， 北京 100080）

[**摘要**]20 年来中文信息处理取得了巨大成绩，这是有目共睹的。当前摆在学界面前的一个重要任务是确立全局的战略目标，并尽快在一些社会急需的发展方向上取得实质性的突破。为此，首先要澄清某些认识， 比如中文信息处理是不是一定要在汉语理解的基础上推进？对于解决中文信息处理的一些急需课题来说，究竟什么方法是最适用的？本文首先对国内外自然语言处理的历史作了一个简短的回顾，说明从小规模受限语言处理走向大规模真实文本处理，是一个不可抗拒的历史潮流。并通过一些具体的实例来说明：统计语言模型能解决什么问题？它为什么在一些有可比评测的课题上连连胜出？藉此阐明，具有统一测试数据和统一计分方法的可比评测是推动科学技术进步的有力杠杆。我们应当拿起这个武器。

[**关键词**]中文信息处理，统计语言模型
[**中图分类法**] H08 [**文献标识码**] A [**文章编号**] 1003-5397(2002)01-0077-08

What Can We Do with Statistical Language Models?

Huang Changning

Abstract: Obviously Chinese information processing (CIP) has got outstanding achievement in the past two decades. The most important task of the community is to establish the strategy objective of CIP, and make essential break-through on certain development directions urgently needed by the society as soon as possible. For the purpose, we want to clarify some remarks first. For example, is it necessary to push forward CIP research based on Chinese language understanding? For those urgently needed CIP projects, what is the most appropriate approach? The paper makes a brief survey on the international history of natural language processing (NLP) first, and points out that the moving from small scale restricted NLP to large scale running text processing is an uncontrollable trend. And then through some concrete examples the paper describes what kind of tasks can be solved by statistical language models (SLM), and why they always outperform than their competitors under comparable evaluations. The comparable evaluation with uniform testing data and scoring method is a powerful lever for achieving progress of science and technology. Let’s arm ourselves with such a weapon.

Key words: Chinese information processing; statistical language model

一 大规模真实文本处理

从 50 年代初机器翻译课题的提出算起，自然语言处理的研发历史至少也有五十年了。了解这段历史的同行，大概都知道我们的研究目标是怎样从小规模受限语言处理走向大规模真实文本处理的。把这个新目标正式列入大会主题的

是 1990 年在赫尔辛基举行的第 13 届国际计算语言学大会（Coling'90）。理由其实很简单，因为那些只有几百词条和数十条语法规则的受限语言分析系统，通常被业内人士戏称为“玩具”，是很难有什么实用价值的。政府、企业和广大计算机用户期盼的是像汉字输入、语音识别、文本检索、信息抽取、信息安全和机器翻译那样的、有能力处理大规模真实文本的实用化系统。当时很多人已经意识到，如果再不思变，这个研究领域是否还有资格存在下去都成了问题。设想一下，如果有一天政府和企业不再资助这类只开花不结果结果的研究，我们还能做什么呢？正是对这段历史的回顾与反思，促使我在 1993 年撰写了《关于处理大规模真实文本的谈话》[1]那篇论文。

那么八年过去了，情况如何呢？还记得当时我在文中列举了下面四种大规模真实文本处理的应用前景：（1）新一代情报检索系统，（2）按客户要求编辑的报纸，（3）从文本到数据库的自动生成，（4）大规模语料库的自动分析。值得庆幸的是，今天所有这四个方向都有了实用化或商品化的成果。第一个任务是信息检索，如网站上普遍使用的搜索引擎。由于电子出版业和因特网的飞速发展，这门技术虽然还有发展空间，但已经相当成熟了。第二个任务现在叫做信息过滤和自动文摘，事实上当前有些报社或公司已经可以提供这样的服务，如《洛杉矶时报》的 MyNews 服务：http://www.latimes.com/services/ 。第三个任务叫信息抽取，虽然国际上也已经有些专门的公司以此营生，如美国的 Cymfony 公司： http://www.cymfony.com/mission.htm ； 但仍有许多技术上的难题没有攻克，比如代词的照应(anaphora resolution), 非受限文笨中的事件识别等等。至于第四个课题，世界各国已建成了多种文字的带标语料库，它们在自然语言处理和语言学研究中发挥了不可替代的作用（试访问北京大学计算语言学研究所的网站：http://icl.pku.edu.cn/introduction/corpustagging.html）。

美国国防部近年启动的 TIDES (Translingual Information Detection, Extraction and Summarization)计划(http://www.darpa.mil/ito/research/tides/)，把语言信息处理研发的一个重点定位在跨语言的信息检索、信息抽取和自动文摘上。实际就是上面提到的前三个任务，再加上机器翻译。但这里所说的并不一定是高质量的全文翻译。因为比如在跨语言文献检索中，最低要求只需有一部双语词典，把查询中的关键词从一种语言翻译成另一种语言（即目标语），然后就可以通过传统的信息检索方法去查找目标语的文档了。

在行将结束这一节的时候应当指出，尽管大规模真实文本处理是一个战略目标，不等于说小规模受限语言处理，如受限领域的机器翻译、语音对话系统、电话翻译系统和其它各种基于深层理解的自然语言分析系统或理论研究，就不应当搞了。目标和任务的多样化也是学术界繁荣昌盛的一个标志。问题是要分清轻重主次，不要自乱阵脚。这对于政府的科研主管部门和研究团体的学术带头人来说尤其重要。

二 方法的争论

目标和方法通常是密不可分的。如果我们同意把大规模真实文本处理作为自然语言处理的战略目标，那么实现这一目标的理论和方法也必然要跟着变化。无独有偶，1992 年在蒙特利尔召开的第 4 届机器翻译的理论和方法国际会议(TMI-92)宣布大会的主题是：“机器翻译中的经验主义和理性主义方法”。公开承认，在传统的基于语言学和人工智能方法的自然语言处理技术以外，还有

一种基于语料库和统计语言模型的新方法正在迅速崛起。方法的多元化无疑是学术界的幸事，问题是为什么在这个时候萌发出这样一种新方法。

其实，做任何研究的人都会遇到方法论的问题。比如做机器翻译，在底层一般有三种方法可供选择：直接法（如早期的 SYSTRAN 系统）、转换法（目前的常规方法）和中间语言法（如 CICC 亚洲五国语言的机器翻译系统等）。在高层则有所谓的经验主义方法（数据驱动、双语语料库、翻译统计模型等）还是理性主义方法（句法-语义分析，中间表示、译文生成等）的争论。

又如，做自然语言的句子分析，一般都会用到词库、句法和语义等知识。但具体实现时，可以是先句法-后语义（常规做法），也可以是句法-语义一体化（如蒙太格语法），还可以直接从语义入手（如 70 年代 Schank[2]、[3]的概念依存理论和概念分析系统等）。了解一下这些方法在历史上的消长，也是有益的。

历史上学术界曾多次在方法论上发生过重大争论。例如，80 年代人工智能界曾对究竟是符号主义还是连接主义展开过热烈争论。讨论的中心是基于神经元网络的连接主义方法会不会最终取代传统的符号方法。Minsky 在 1990 年的一篇著名的论文[4]中对此做出了否定性的回答。他认为人工智能不同于电路理论和电磁学。在这里不可能有像电路理论中的克希荷夫定律，或电磁学中的马克斯韦尔方程那样神奇的统一理论。他主张人工智能必须利用各式各样的方法，包括各种不同的知识表示。他相信采用多元化的构件来建造复杂的人工智能系统的时候已经来到，这些构件中有的是连接主义的，有的是符号主义的，每个构件都有它自身存在的理由。

在自然语言处理系统中，我们都承认语言知识是不可或缺的。但是，如此海量的知识究竟是靠语言学家的直觉或语感来获取呢？还是从语料库大量可观察的语言事实中归纳出来？这就是 90 年代初那场理性主义和经验主义讨论的焦点。如果自然语言处理的目标是大规模真实文本处理，那么经验主义就成了首选的方法。因为除此之外我们实在不知道还有什么方法可以用来获取如此浩瀚的语言知识。今天即使是语言学家也必须利用语料库提供的证据和实例。Quirk 等编著的《英语语法大全》[5]利用了语料库的数据。近年来国外著名出版社编撰的英语词典，几乎没有一部不是在语料库的支持下完成的。

其实每一个研究人员都不可能回避这样或那样的方法论问题。不同的任务应当选择不同的方法。对某一种特定的任务来说，哪种方法更好其实正是学术交流的一个中心话题。而高低优劣是可以通过评测来判断的。下面就让我们具体看一看统计语言模型（即经验主义方法）是怎样解决语言信息处理问题的， 它们又在哪些任务上取得了进展。

三　统计语言模型

为了阐明经验主义方法对大规模真实文本处理的贡献，这一节将通过一些大家熟悉的实例, 一方面具体地看看统计语言模型是怎么工作的，另一方面用评测结果来说明概率统计方法的优势。受篇幅的限制，本文将主要介绍常用的 N 元模型。

1. N 元模型

如果用变量 W 代表一个文本中顺序排列的 n 个词，即 $W = w_1w_2…w_n$，则统计语言模型的任务是给出任意词序列 W 在文本中出现的概率 $P(W)$。利用概率的乘积公式，$P(W)$可展开为：

$$P(W) = P(w_1)P(w_2|w_1)P(w_3| w_1 w_2)…P(w_n|w_1 w_2…w_{n-1})$$

不难看出，为了预测词 w_n 的出现概率，必须已知它前面所有词的出现概率。从计算上来看，这太复杂了。如果任意一个词 w_i 的出现概率只同它前面的 *N-1* 个词有关，问题就可以得到很大的简化。 这时的语言模型叫做 N 元模型 (N-gram)，即

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|\ w_1\ w_2)\ldots P(w_i|w_{i-N+1}\ldots w_{i-1})\ldots$$
$$\approx \prod_{i=1,\ldots,n} P(w_i|w_{i-N+1}\ldots w_{i-1})$$

符号 $\prod_{i=1,\ldots,n} P(\ldots)$ 表示概率的连乘。实际使用的通常是 N=2 或 N=3 的二元模型(bi-gram)或三元模型(tri-gram)。以三元模型为例，近似认为任意词 w_i 的出现概率只同它紧前面的两个词有关，即

$$P(W) \approx \prod_{i=1,\ldots,n} P(w_i|w_{i-2}\,w_{-1})$$

重要的是这些概率参数都是可以通过大规模语料库来估值的。比如三元概率有

$$P(w_i|_{wi\text{-}2wi\text{-}1}) \approx count(w_{i-2}\,w_{i-1}\,w_i) / count(w_{i-2}\,w_{i-1})$$

式中 *count(…)* 表示一个特定词序列在整个语料库中出现的累计次数。

统计语言模型有点像天气预报的方法。用来估计概率参数的大规模语料库好比是一个地区历年积累起来的气象纪录，而用三元模型来做天气预报，就像是根据前两天的天气情况来预测今天的天气。天气预报当然不可能百分之百正确。这也算是概率统计方法的一个特点吧。

2. 语音识别

语音识别作为计算机汉字输入的另一种方式越来越受到业内人士的青睐。所谓听写机就是这样的商品。据报道中国的移动电话用户已超过一亿，随着移动电话和个人数字助理（PDA）的普及，尤其是当这些随身携带的器件可以无线上网的时候，广大用户更迫切期望通过语音识别而不是小键盘来输入简短的文字信息。无怪乎近年来语音识别的研发成了国际上的大热门。那么当前商品化的听写机又采用什么技术呢？是不是像许多人想象的那样必须建立在汉语理解的基础上呢？

其实，语音识别任务可视为计算以下条件概率的极大值问题：

$$W^* = argmax_W\, P(W|speech\ signal)$$
$$= argmax_W\, P(speech\ signal|W)\ P(W) / P(speech\ signal)$$
$$= argmax_W\, P(speech\ signal|W)\ P(W)$$

式中数学符号 $argmax_W$ 表示对不同的候选词序列 *W* 计算条件概率 *P(W|speech signal)*的值，从而使 *W**成为条件概率值最大的那个词序列。它也就是当前输入语音信号 *speech signal* 所对应的输出词串。

公式第二行是利用贝叶斯定律转写的结果，因为条件概率 *P(speech signal|W)* 比较容易估值。公式的分母 *P(speech signa)* 对给定的语音信号是一个常数，不影响极大值的计算，故可以从公式中删除。在第三行所示结果中，*P(W)*叫做统计语言模型；*P(speech signal|W)* 叫做声学模型。

讲到这儿，细心的读者可能已经明白，汉语拼音输入法中的拼音－汉字转换任务其实也可以采用同样的方法来实现，而且两者所用的汉语语言模型（即三元模型）是同一个模型。

据调查，目前市场上的听写机产品和微软拼音输入法（3.0 版）都是用词的三元模型实现的, 几乎完全不用句法-语义分析手段。为什么会出现这样的局面呢？我看也是性能评测所为，优胜劣汰嘛。据我们所知，用三元模型实现的拼音-汉字转换系统是目前市场上性能最好的产品。可比的评测结果表明，它的出错率比其它产品减少约 50%。

应当指出，三元模型（或 N 元模型）其实只利用了语言的表层信息（或知识），即符号（字、词、词性标记等）序列的同现信息。谁也没有说它是完美的。在这一领域中，下一个研究目标应当是结构化对象的统计模型。当然能做到语言理解是了不起的成果，它肯定会比目前这种统计语言模型强得多，这是不争的事实。问题是目前国内外还没有哪一种句法-语义分析系统可以胜任大规模真实文本处理的重任。

3. 词性标注

至少像短语结构文法这样一类的语法规则是建立在词类基础上的。无怪乎语言学界有句行话说，没有词类就没法讲语法了。所以在自然语言的句法分析过程中，大概都有一个词性标注的阶段。不难理解，汉语的自动分词和词性标注的精确率，将直接影响到后续的句法分析结果。据观察，在汉语句句法分析结果中，有高达 60%的分析错误来源于分词和词性标注的错误。

在英语的词库中约 14% 的词形(types)具有不只一个词性，而在一个语料库中，总词次数(tockens)的约 30% 是兼类词。从这个统计数字中可以估计出词性标注任务的难度。历史上曾经先后出现过两个方法迥异的英语词性标注系统：TAGGIT 系统拥有 3000 条上下文相关规则， 而 CLAWS 系统[6]完全采用概率统计方法。两个系统各自完成了 100 万词次的英语语料库的自动词性标注任务。评则结果（见下表）表明，采用概率统计方法的 CLAWS 系统的标注精度达到 96%，比 TAGGIT 系统提高了近 20 个百分点。我本人对用概率统计方法来处理大规模真实文本的认识，正是从当年这个鲜明对比中形成的。

系统名	TAGGIT(1971)	CLAWS(1987)
词类标记数	86	133
方法	3000 条规则	二元模型
标注精度	77%	96%
测试语料库	布朗语料库	LOB 语料库

具体来说，CLAWS 系统采用的是词类标记的二元模型。如果令 $C = c_1 \ldots c_n$ 和 $W = w_1 \ldots w_n$ 分别代表词类标记序列和词序列，则词性标注任务可视为在已知词序列 W 的情况下，计算如下条件概率极大值的问题：

$$C^* = \mathrm{argmax}_C P(C|W)$$
$$= \mathrm{argmax}_C P(W|C)P(C) / P(W)$$
$$\approx \mathrm{argmax}_C \prod_{i=1,\ldots,n} P(w_i|c_i)P(c_i|c_{i-1})$$

$P(C|W)$ 表示：已知输入词序列 W 的情况下，出现词类标记序列 C 的条件概率。数学符号 argmax_C 表示通过考察不同的候选词类标记序列 C, 来寻找使条件概率 $P(C|W)$ 取最大值的那个词序列 W^*。后者应当就是对 W 的词性标注结果。

公式第二行是利用贝叶斯定律转写的结果，由于分母 $P(W)$ 对给定的 W 是一个常数，不影响极大值的计算，故可以从公式中删除。接着对公式进行近似。首先，引入独立性假设，认为词序列中的任意一个词 w_i 的出现概率近似只同当前词的词性标记 c_i 有关，而与周围（上下文）的词类标记无关。即词汇概率

$$P(W|C) \approx \prod_{i=1,\ldots,n} P(w_i|c_i)$$

其次，采用二元假设，即近似认为任意词类标记 c_i 的出现概率只同它紧邻的前一个词类标记 c_{i-1} 有关。有

$$P(C) \approx \prod_{i=1,\ldots,n} P(c_i|c_{i-1})$$

$P(c_i|c_{i-1})$ 是词类标记的转移概率，也叫做二元模型。

上述这两个概率参数都可以通过带词性标记的语料库来分别估计：

$P(w_i|c_i) \approx count(w_i,c_i) / count(c_i)$

$P(c_i|c_{i-1}) \approx count(c_{i-1}c_i) / count(c_{i-1})$

顺便指出，国内外学者用词类标记的二元或三元模型实现的汉语文本词性自动标注也同样达到了约 95%的标注精确率[7]。

4. 英语介词短语消歧

下面介绍的是另一种概率统计方法。在形式上它不同于 N 元模型。但从带标的语料库中发掘以概率形式表示的语言知识，两种方法在本质上又十分相似。

众所周知，介词短语附加（PP attachment）是英语句法分析中著名的结构歧义问题。以下面的例句来看:

Pierre Vinken, 61 years old, joined the board as a nonexecutive director.

介词短语“as a nonexecutive director”可能修饰前面的动词“joined”，也可能修饰前面的名词短语“the board”（指机器而言，对人来说并没有歧义）。为了用计算机来求解介词短语的正确附加，就先要找到一种合适的知识表示（即形式化）来描述这个问题。比如，用 *A*=1 表示名词附加，*A*=0 表示动词附加。接着，分别用句中动词短语的中心词 *V*、作为该动词宾语的名词短语的中心词 *N1*、介词 *P* 和作为介词宾语的名词短语的中心词 *N2*，来表示输入句子的骨架。由于例句中的介词短语是修饰动词“joined”的，所以它的表达式如下：

(*A=0, V=joined, N1=board, P=as, N2=director*)

有了形式化表示之后，如果你还有一个已经标注好句法结构的语料库，如宾州数库（UPenn Treebank）。你就可以计算上述每个个别表达式（四元组）的出现次数，进而统计出它出现的条件概率 *Pr (A=1 | V=v, N1=n1, P=p, N2=n2)*。当然，以个别词而不是词类来进行统计，会出现严重的数据稀疏问题（即训练不足）。所以在具体的实现中，可以采用所谓的后撤（back-off）算法。当找不到四个中心词的四元组时，就退一步找三个中心词的三元组，以此类推，直至退到一元组时，只根据具体的介词来作出判断。

判断介词短语附加的具体过程如下：

若 *Pr (1 | v, n1, p, n2) ≥ 0.5*，

则 判定介词短语 PP 附加于 *n1*；

否则 附加于 *v*。

下面是 Collins 和 Brooks 的论文[8]中用这种概率统计方法得到的实验结果。他们的评测是这样进行的：语料库采用美国宾州大学提供的带有句法表注的 华尔街日报（WSJ）树库，从中抽出 20,801 个四元组作为训练集，其余的 3,097 个四元组作为测试集。

饶有兴趣的是他们通过实验，得出了介词短语自动判定的上、下限。实验根据如下：

实验条件	精确率
一律视为名词附加 (即 A≡1)	59.0%
只考虑句中介词 p 的最常见附加	72.2%
三位专家只根据四个中心词判断	88.2%
三位专家根据全句判断	93.2%

很明显，自动判断精确率的下限是 72.2%，因为机器不会比只考虑句中介词 p 的最常见附加做得更差了；上限是 88.2%，因为机器不可能比三位专家根据四个中心词作出的判断更高明。

论文报告，在被测试的 3,097 个四元组中，系统正确判断的四元组为

2,606 个，因此平均精确率为 84.1%。这与上面提到的上限值 88.2%相比，应该说是相当不错的结果。

四 可比评测是唯一的评判标准

有评测才会有鉴别。评判一种方法优劣的唯一标准是相互可比的评测，而不是设计人员自己设计的“评测”，更不是人们的直觉或某个人的“远见”。近年来，在语言信息处理领域，通过评测来推动科学技术进步的范例很多。国家 863 计划智能计算机专家组，曾对语音识别、汉字（印刷体和手写体）识别、文本自动分词、词性自动标注、自动文摘和机器翻译译文质量等课题进行过多次有统一测试数据和统一计分方法的全国性评测，对促进这些领域的技术进步发挥了非常积极的作用。但是这期间也遇到了一些阻力，有些人试图用各种理由来抵制这样的统一评测，千方百计用“自评”来取代统评。其实，废除了统一的评测，就等于丧失了可比的基础。这个损失使得上述任何理由都变得异常苍白。在这里我想对我国研究生说几句话：统一的、可比的评测是科研工作的一条基本原则。研究生的论文想被顶尖的学术刊物或会议所接受，最好遵循这条原则。

在国际上，美国国防部先后发起的 TIPSTER 和 TIDES 计划，干脆叫做“评测驱动的计划”。该计划在信息检索(TREC：http://trec.nist.gov/)、信息抽取（MUC）和命名实体识别（MET-2）等研究课题上，既提供大规模的、统一的训练语料和测试语料，又提供统一的计分方法和评测软件。以此保证每个参加评测的研究小组都能在一种公平、公开的条件下来进行方法的比较，追求科学技术的进步。TREC, MUC 和 NET-2 等这些系列会议要求与会小组，必须像奥运会那样在会前提交统评结果，才能到会上来报告各自的方法和相应结果。它的优点是可以完全摒弃那些假行家，使他们的大话空话不再有市场。1998 年 TREC 在它的第 7 届年会上宣布，经过与会各方七年来的努力，美国信息检索的效率(effectiveness)提高了约一倍。此外，英、中、日文的命名实体(NE)识别，经过类似的评测机制，其召回率和精确率也都分别达到了实用化的水平。这些组织完善的计划一方面鼓励多种方法的试验，另一方面通过统一的评测对每种方法作出公平的裁决，使得好方法能及时脱颖而出。

为了推动中文信息处理的发展，让我们拿起评测这个武器。建议政府科研主管部门在制定项目计划时，至少要在一个项目的总经费中拿出 10%的拨款用于资助该项目的评测。这不仅因为评测本身就是一个重要的科研课题，需要有很大的投入，而且因为没有统一评测的研究项目，其成果毕竟是不可信的。

[参考文献]

[1]黄昌宁：关于大规模真实文本处理的谈话，《语言文字应用》1993 年第 3 期。

[2]Schank, R., and Abelson, R. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale: Lawrence Erlbaum Associates, Publishers,1977.

[3]Rich, Elaine. *Artificial Intelligence*. pp.295-344, London: McGraw-Hill Book Company, 1983.

[4]In: *Artificial Intelligence at MIT: Expending Frontiers*, Vol.1. Winston, P. H., and Shellard, S.A. (eds.). Cambridge, Mass: MIT Press, 1990
[5]Garside, R., Leech, G. and Sampson, G. (eds.). *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman, 1989
[6]夸克等《英语语法大全》，华东师范大学出版社，1988
[7]白拴虎《汉语词性自动标注系统研究》，清华大学计算机科学与技术系硕士学位论文，1992
[8]Collins, M. and Brooks, J. Preposition phrase attachment through a backed-off model. In: *Proceedings of the 3rd Workshop on Very Large Corpora (WVLC)*, Cambridge, MA, 1995