KDD和基于语料库的机译系统

范建华　李德毅　张亚非
（南京通信工程学院计算机系，210016）

摘　要： 知识获取是人工智能研究的中心问题，但是事实证明从领域专家那里获取知识具有很大的局限性，因此计算语言学目前开始主要从语料库中获取语言知识。KDD（Kowledge Discovery in Databases）是专门研究从大量数据中发现隐含知识的一门新技术。本文初步探讨了如何将它应用于一个基于语料库的机器翻译系统中。

KDD&Corpus-Based Machine Translation

Fan Jianhua　Li Deyi　Zhang Yafei
(Nanjing Communication Engineering Institute, 210016)

Abstract： How to acquire knowledge is one of the most significant tasks in Artificial Intelligence Systems. In current Computational Linguistics, corpus is one of the main sources from which language knowledge is acquired. KDD is a new technology which aims to find out knowledge from databases . This paper discusses how this technology is applied to a corpus-based machine translation system.

一、KDD和发现状态空间理论

KDD一词首次出现在1989年8月第十一届国际人工智能会议上。概括地说，KDD是把数据库作为知识源，综合运用逻辑学、统计学、机器学习、模糊学、数据分析、可视化计算等，从中发现隐藏在大量数据间的知识。

随着计算机应用的普及和数据处理在计算机应用中所占比重的上升，数据库技术得到了迅猛的发展。随着数据库容量的增大，简单的数据查询方式已经很难满足某些信息系统的要求。因此从数据库中发现知识已经成为当前一个极具挑战性的研究课题。【1】

经过几年不懈的努力，该领域取得了很大进展。目前，国际上一些KDD的研究成果已在某些领域得到应用，例如美国信息资源公司和MIT的John D.C. Little合作开发的一个商业系统－CoverStory和美国著名KDD专家G.Piatetsky Shapiro等人开发的大型数据库交互分析工具系统Knowledge Discovery Workbench (KDW)等等。在国内，李德毅教授领导的研究小组对KDD进行了深入细致的研究，创立了发现状态空间理论，并先后在 DOS和 Windows平台上开发出了KDD原型样机。

下面简单介绍一下发现状态空间理论。

发现状态空间是发现系统实施多种发现算法的运作空间。首先把原始数据库中和发现任务相关的所有数据汇集在一起，形成知识基，也是最初的知识模板；然后，对基底进行

统计运算，形成宏元组。在发现状态空间内进行的多种知识汇集操作分成三个方向，即面向属性的操作，面向宏元组的操作，从微观到宏观的操作。依靠由知识模板决定的知识熵和发现难度所构成的目标函数引导、反复进行这些操作，就能得到所需的知识。其主要的手段是归纳、类比、联想、证伪和演绎等。

二、机器翻译和语料库语言学

机器翻译是自然语言处理中最活跃的领域之一，早在五十年代就有人开始研究。70年代以前，其主要的理论依据是Chomsky的生成语法理论，语言分析的主要工具是句法规则。但是由于自然语言本身是个极其复杂的系统，它拥有丰富多彩的语言形式，能够表达极其复杂微妙的语义内容，而规则很难描述所有的语言现象，因此基于规则的机器翻译难以取得突破性的进展。

语料库语学是70年代兴起的另一种语言学方法，它主要针对了规则模型的困境，提出以语料库和统计模型为基础，从语料库中存储的大规模真实文本中直接获取多种语言知识。由于采用这种方法获得的语言知识覆盖面宽，因此能大大改进语言处理系统的处理质量。目前，语料库语言学的发展极为迅速，人们根据不同的应用目的建立不同的语料库。在机器翻译领域，语料库的使用主要存在两种方法：纯统计法和例子模仿法。

- **纯统计法。**

它由Warren Weaver于1949年首先提出，并成功地用在了语音识别系统中。后来，P.Brown等人继承和改进了Weaver的思想，并把它用于机器翻译领域。其基本思想是：一种语言的一个句子是另一种语言中任何句子的一种可能对应。对于每对句子（S，T）（S表示原语言句子，T表示目标语言句子）都赋于一个概率Pr（T|S），表示当原语言中S出现时，人们译出T的可能性。机器翻译可看作是这样的一个过程：给定一个目标语言句子T，去搜索人译出T的原句子S。由Bayes定理：

$$Pr(S|T) = Pr(S) \times Pr(T|S) / Pr(T)$$

Pr（S）称为S的语言概率，Pr（T|S）称为T在给定S下的翻译概率。因此，一个统计MT系统需要一个计算语言概率的模型和一个计算翻译概率的模型，以及寻找使Pr(S）× Pr(T|S）最大的原语言句子的方法。模型的参数是从一个大型双语语料库中用统计方法来估测出来的。

- **基于例子的方法**

基于例子的机译方法是1984年由日本的Nagao博士首先提出的，当时他称为“Memory-based Translation”。基原理是：建立一个语言数据库，其中存放大量的例子及其对应的译文（例子可以是各种语言成分，包括词、片语和句子，甚至篇章），在翻译时，系统从语料库中抽取与输入成分类似的例子，然后模仿例子来完成从原语言到目标语言的转换。由于这种方法不考虑或很少考虑语法语义，有人称之为“野蛮翻译”。

三、KDD在机器翻译中的应用

根据语料库语言学和KDD技术的发展背景我们都可以看出，传统的知识获取方法已经不适合当前信息处理的要求，大规模的的真实数据才是知识的真正源泉。由于语料库语言学是单纯从自然语言处理的角度来组织和利用语料库的，而KDD的理论和技术是从一个相对比较抽象的层次来分析如何从大量数据中发现知识的，因此我们尝试把KDD的技术应用于语料库语言学，具体地说，就是把KDD技术应用在一个基于语料库的面向句子的英汉/汉英双向机器翻译系统中。

系统拟包括语料句库、机器词典、规则库等，翻译过程由特征提取，知识获取，例子匹配，翻译调整等组成。

1、 语料库的构造

我们选取英语900句作为主要的分析语料，其理由是这些句子是语言专家在语言实践中经过精心搜集，然后归纳挑选出来的。其中的句子都具有很强的代表性和实用性，内容也非常丰富，所以很适合作为一个模型系统的语料。

语料库是整个系统的基石，它的结构决定了整个系统知识发现（获取）的能力和效率。要想从语料库中真正获取语言知识，就必须对库存语料进行词法、句法、语义等各种层次上的加工，这一步骤使得语料由“生”变“熟”，这样才能使知识获取成为可能。关于语料加工的理论、方法和工具在目前计算语言学界中存在许多争议。【9】针对要在语料库中实施KDD方法的这个特点，我们对语料的加工主要有以下几个方面：分词标注、词性标注、句子结构标注、对应词标注。

关于汉语分词标注的规范我们主要根据汉语词汇用法的统计结果，使用频度高的汉字组合【10】。词类划分主要包括：实动词、名词、形容词、副词、数词、介词、助动词、虚词、其它等。这主要是因为在英语和汉语中，这些词之间存在较强的对应关系，而且也符合人分析句子的常规，因而可操作性比较强，方便人工进行标注。关于句子结构的标注（即动词类型标注），主要依据牛津字典中英语动词的分类方法（把动词划分成25种类型）。对应关系的标注包括句子在各个层次上的对应关系，如单词层面，习惯用语层面，词组层面，甚至句子层面。如果输入句子能在句子这一层面上找到对应的目标句子，那么在翻译时就不必要进行句子内部语言成份的分析，直接查询就可以得到翻译结果。

2、 特征提取

特征抽取是对句子进行初步的分析，获取那些根据句子表层现象得到的句子特征，这些特征包括：句长（包括基于字母的长度和基于单词的长度）、关键词（如“是”，“将”，“be”、“will”，“when”等）、标点符号等等。

3、 知识获取和例子匹配

根据抽取出来的特征，回到语料库，从中发现该类句子的翻译规则（知识）。KDD的方法在这里将得到充分的利用，根据具体的例子，归纳、类比、联想、演绎等手段都将要不同程度地使用。而特别重视归纳方法，是该系统的一个显著特色。因为，语言现象是一种随机现象，语言学的规则大多数都是随机性的规则，因此，在翻译过程中有必要采用统

计的方法，以提高翻译的效率。【11】有了统计的结果，我们就可以总结归纳出有意义的语言转换规则。语言统计可以在不同层次上进行，如英语就有字母统计、词频统计、单词同现频率统计、句子语法结构统计、句子长度统计等，汉语有字频统计、词频统计、词的组合频率统计、句长统计等。这些统计结果在归纳过程中可以起到不同程度的作用，为此，需要建立相应的频率词典。例如，英语句子语法结构的统计结果和汉语词的组合频率结果，在一词多义或一词多性的情况下就可以有效地帮助解歧。例如在句子：A light tap sounded on the door.（门上响起了轻轻的敲门声。）我们给出各个词可能的词性(A为冠词，N为名词，V为动词，P为介词，A为形容词)：

A light tap sounded on the door

D N N V P D N

A V

V

前三个词可能的语法组合是6种，但是如果在频率词典中查出DAN的同现频率远远大于其它类型的组合，那么我们就可以初步判定单词light是形容词，tap是名词。这样可以大大提高分析的效率。

此外，归纳还体现在另一个更为重要的方面，那就是匹配过程中的概念提升。我们知道，结构相同或相似的句子其译文结构也大致相同或相似。于是我们把句子结构归结为若干类（英语句子可以参见牛津英汉双解词典所划分的二十五类，汉语句子有待于进一步研究），每一类都称作一个模板（template），例如：

It is important to do this job. 和 It is impossible to go further.

这两个句子具有相同的模板：“it+BE+subject.”

我们认为，具有相同模板的句子其译文的模板也相同。这样，在匹配以前，就必须把输入句子的模板抽象出来。匹配时不但要进行句子匹配，更重要是进行模板匹配。由于句子中可能存在多个可以成为谓语动词的单词，而且每个单词又可能有多个义项，因此抽取出来的模板就有多种可能。这种情况可以用上述统计的方法来全部解决或部分解决。如果不能完全解除歧义，就把所有的可能都输出，在译文转换阶段来进一步判断。所以抽取出来的模板还是有待验证的。

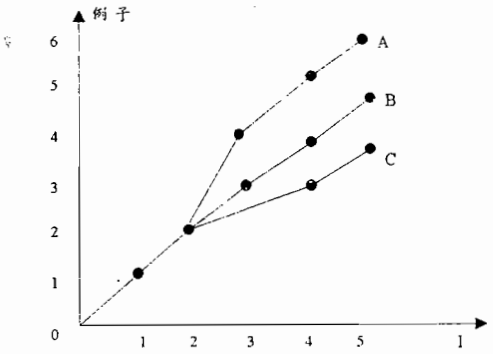

图1 基于单词层面的简单句子匹配

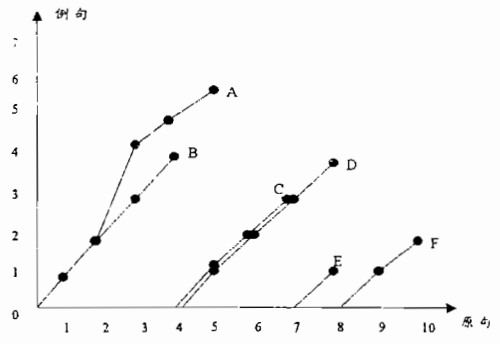

图2 多个例句覆盖一个原句示意图

除了模板匹配外，在单词层面上也可以进行匹配。如例子A、B和C，输入句子是I:

A：abfcde　　B：abcde　　C：abde　　I：abcd

因此就有如图1所示的结果。

当然，图1所示的是存在一个例子能完全复盖原句的几种情况，但是，更多的情况是只有两个或三个以上的例句才能完全复盖原句。如图2。这时，就必须把这些例句都放入候选队列，待译文转换时加以处理。

4、译文转换

抽取出例子后，我们可以得到一个候选例子集及其各种标注信息。这一阶段的主要任务是用抽取出来的例子覆盖原句，比较原句和例子之间的不同点，形成一个转换表。最后根据转换表用目标语言成分进行翻译重组，形成译文。这里要分两种情况:

A、单个范例覆盖原句的转换

首先选一个相似程度最高的例子。为此我们必须计算句子之间的相似程度。句子之间的相似程度一般是具有相同模板的前提下通过可替换单词或词组之间的相似度来计算的，对于不同的单词我们定义它们之间的相似度，根据相似度的高低来选取不同的例句。一般来说，同词性的词之间比词性不同的词之间具有更高的相似度，同类的词之间比不同类的词之间具有更高的相似度。如“friend”和“guest”这两个词之间的相似度就比“friend”和“letter"之间的要高，例如，现有原句“I received a guest yesterday.”根据它在语料库中抽取出这样两个例句:

I received a friend yesterday.　昨天我接待了一位朋友。

I received a letter yesterday.　昨天我收到了一封信。

这两个例句具有相同的模板，与原句都很类似，但是由于“friend”和“guest”有更高的相似度，所以最后选择“接待”作为“receive”的译文。因此，用相似度比较的方法能较好地解决前一步骤遗留下来的问题，使具有相同模板的句子歧义现象得以解决。选择了相似度最高的例子后，就可以比较例句和原句之间的差别，形成一个转换表。这种差别主要在单词层面，也可以在词组层面。如下表:

	相似度	句子
原句	0	他明天要去南京。
例句1	5	我明天要去北京。
例句2	4	他下午要去图书馆。

转换表

原文	译文
我→他	I→he
北京→南京	Beijing→Nanjing

最后根据转换表，生成译文：He will go to Nanjing tomorrow.

B、多个例句覆盖原句的转换

当有多个范例才能覆盖整个原句时，就必须把这些范例联结起来。显然，我们很容易碰上同时有多种组合都能覆盖原句，因此又必须选择那种能覆盖原句而且“耗费”又是最小的组合。各段根据各自的转换表进行转换，最后把转换后的各段进行联结，形成译文。

5、译文调整

译文调整是一个非常必要的过程，这是因为即使是很相似的句子其译文也未必完全类似，必须根据语言习惯和最基本的语言规则来进行验证和调整。为此，我们将利用传统语言学的规则来辅助翻译。以往实践证明，译文调整对提高译文质量有不小的贡献。

四、结语

本文介绍了KDD技术及其在一个基于语料库的机器翻译系统中的应用，其知识获取的主要方法是归纳、类比、联想和演绎等。我们进一步研究的问题是发现状态空间理论如何在一个标注的语料库中应用。目前，该系统的规模还很小（语料库只有500个句子），而且没有收入复句，因此还有许多工作有待进一步研究。

参考文献

【1】李德毅，发现状态空间理论，小型微型计算机系统，No.11，1994.

【2】Nagao, M, A Frame work of a Mechanical Translation between Japanese and Engish by Analogy Principle, in Artificial and Human Intelligence, eds A. Elithornand R. Banerji, North-Holland, 1984.

【3】P.F.Brown, et, al: A Statistical Approach to Machine Translation, Computational Linguistics, 1990, vol. 16, No.2.

【4】Hiroaki Kitano, A Comprehensive and Practical Model of Memory-Based Machine Translation, Proceedings of IJCAI-93, 1993.

【5】李德毅，知识获取和数据库学习系统，计算机科学，Vol.20.No.5，1993.

【6】Sato.S, Nagao. M, Toward Memory-Based Translation, Procedings of Coling-90, 1990

【7】Sumita. E, Example-Based Machine Translation on Massively Parallel Processors, Proceedings of IJCAI-93, 1993

【8】周明 黄昌宁，面向语料库标注的汉语依存体系的探讨，中文信息学报，Vol.8.No.3.1994

【9】汉语词汇的统计和分析，外语教学与研究出版社，1985.4

【10】冯志伟 杨平，自动翻译，上海知识出版社，1987.11,P207-213

【11】张敏 罗振声，语料库和知识获取模型，中文信息学报，Vol.8No.1，1994.1.