DOI:10.13458/j.cnki.flatt.000242

从语料库中挖掘知识和抽取信息

冯志伟
(教育部语言文字应用研究所,北京 100010)

摘 要:本文主要介绍中国传媒大学计算语言学博士生们从依存树库中获取语言学知识的一些工作,如汉语复杂网络的研究等。这些工作都是在汉语依存树库的基础上进行的。本文也简要地介绍了国外从语料库中获取非语言学知识的研究以及信息自动抽取的研究。作者认为从语料库挖掘知识和抽取信息应当是现代语言学研究的基本方法。

关键词:语料库;依存树库;知识挖掘;信息抽取

中图分类号:H0 文献标识码:A 文章编号:1004-6038(2010)04-0001-07

大规模的真实文本语料库包含着无比丰富的知识和信息。语料库是一个宝藏,从语料库中可以挖掘的知识,可以是语言学的知识,也可以是非语言学的其他有用的知识,从语料库中还可以抽取各种各样的信息。

本文首先讨论怎样从语料库中挖掘语言学知识,然后再讨论怎样从语料库中挖掘非语言学知识,最后介绍怎样从语料库中抽取信息。

1 从语料库中挖掘语言学知识

语言学的研究必须以语言事实作为根据,必须详尽地、大量地占有材料,才有可能在理论上得出比较可靠的结论。传统的语言材料的搜集、整理和加工完全是靠手工进行的,这是一种枯燥无味、费力费时的工作。计算机出现后,人们可以把这些工作交给计算机去作,大大地减轻了人们的劳动。后来,在这种工作中逐渐创造了一些独特的方法,提出了一些初步的理论,形成了一门新的学科——语料库语言学(corpus linguistics),由于语料库是建立在计算机上的,因此,很多学者把它看成是自然语言处理(natural language processing)的一个分支学科。

语料库究竟有什么用处。这里我们通过一个实例来说明。

如有关副词“多半”用法的例句:

①游览北京名胜古迹的多半是外地人。(表示“大部分”)

②过了立秋,天气多半会变得凉爽起来。(表示“通常”)

③他们多半会同意的,你不用着急。(表示“很有可能”)

仔细观察,发现句 3有歧义。除了表示“很有可能”之外,还可以表示“他们”中的“大部分”。也就是说,“多半”的语义指向可以向后指向“同意”,还可以向前指向“他们”。

最近,中国传媒大学计算语言学博士生高松带着这样的问题,对北大语料库提供的 500条语料进行分析,得出了如下的统计结果:

	条目数	比例
切分错误	22	4.4%
无歧义	329	65.8%
有歧义	149	29.8%
合计	500	100%

她还发现,如果不分词,会产生如下的切分错误句子:

④我差不多半年都没去书店了。

其实句子 4中根本没有“多半”这个单词。

在有歧义的 149条中,歧义格式可以分为两类:

——名词、名词性短语 +多半 +动词

⑤考到外地大学生又多半不想回来。

——人称代词 +多半 +动词

⑥她们多半是妙龄女子。

高松进一步分析发现,出现歧义的条件是:句子的主语必须是群体性的名词、名词词组或者人称代词。

句 3之所以有歧义,就是因为主语“他们”是表示群体的人称代词。这样就解释了句 3出现歧义的原因。这样的解释是前辈语言学家没有做到的。

高松发现了前辈语言学家没有观察到的问题,做到了前辈语言学家做不到的事情,语料库给她提供的一种观察语言现象的手段,使她有可能获取到重要的语言学知识。可见,语料库确实是语言研究的有力工具,语料库可以帮助普通的年轻学子超越前人。

中国传媒大学计算语言学博士生们近年来在从语料库中获取语言学知识方面做了一些初步的探索,我们不仅使用普通的语料库来获取知识,还进一步把语料库加工成树库(tree-bank)来获取知识。

我们从语料库中获取知识的过程大致如下:

语料库数据→带标语料库→树库→数据挖掘→结构化的数据→统计分析→知识(包括语言学知识和非语言学知识)。

刘海涛和胡凤国把依存树库中的依存树转换成汉语依存网络,使用“复杂网络”(complex network)的理论和方法对依存网络进行了研究。

例如,英语句子“The student has a book”和“He reads an interesting book”的依存树库如图 1。

把树库中的结点加以合并,形成图 2的依存网络(右图

作者简介:冯志伟,研究员,博士生导师,研究方向:计算语言学

是以单词为结点的依存网络,左图是以词类标记为结点的依存网络)。

Annotatopn of two English sentences in the treebank

Order number of sentence	Dependent			Governor			Dependency type
	Order number	Word	POS	Order number	Word	POS	
S1	1	the	det	2	student	n	atr
S1	2	student	n	3	has	v	subj
S1	3	has	v				
S1	4	a	det	5	book	n	atr
S1	5	book	n	3	has	v	obj
S2	1	he	pr	2	reads	v	subj
S2	2	reads	v				
S2	3	an	det	5	book	n	atr
S2	4	interesting	adj	5	book	n	atr
S2	5	book	n	2	reads	v	obj

图 1 英语的依存树库

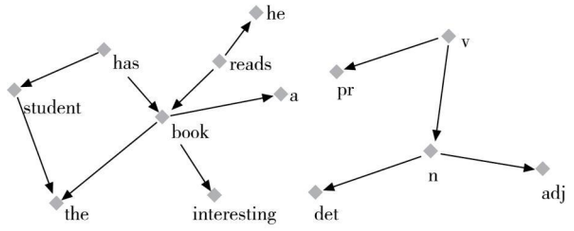

图 2 依存句法网络

使用这样的方法,对于汉语的《新闻联播》(xwlb)树库中的单词结点进行合并,形成了如下的汉语句法网络:

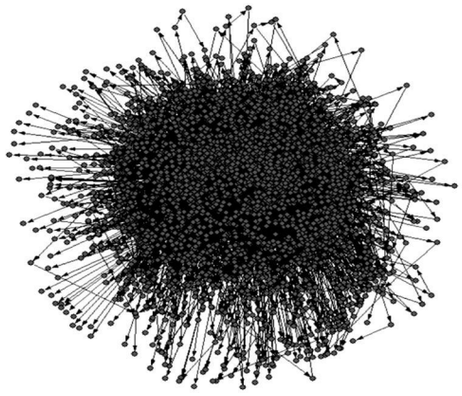

图 3 汉语《新闻联播》树库形成的汉语依存句法网络

他们使用复杂网络理论对这个依存句法网络进行分析。

为了衡量一个网络的复杂性,最常用的复杂网络参数是平均路径长度、聚集系数和度分布。我们围绕这三个参数对句法网络进行考察研究。

—— 句法网络的平均路径长度表示的是网络中任意两个结点之间的平均最短路径长度,用 <d>表示。

—— 在句法网络中,结点的度 k指的是与该结点相连的其他结点的数目 (或边数),它在一定程度上反映了一个结点在网络中的重要性。全部结点的度的平均值称为句法网络的平均度,它反映了句法网络中词与词之间的平均组合能力。结点的度分布通常用分布函数 P(k)描述,它表示一个随机选定词的度恰好为 k的概率。

—— 聚集系数 C是一种用来衡量网络聚集倾向或小集群形态的指标,它度量的是句法网络中一个结点的两个相邻结点之间互连的可能性。

如果一个网络有较小的平均路径长度 d和较大的聚集系数 C,那么,这种网络是一种“小世界”(SmallWorld)网络。汉语句法网络的节点数为 4017,平均路径长度较小,<d>为 3.372,聚集系数 C较大,为 0.260,所以,汉语句法网络可算是一种小世界网络。

1967年,美国社会心理学家 Stanley Milgram (米尔格兰姆, 1933—1984)曾经想要描绘一个连结人与社区的人际联系网,提出了“六度分隔”(Six Degrees of Separation)理论。这个理论可以通俗地阐述为:“任何两个陌生人之间所间隔的人不会超过六个,也就是说,最多通过六个人,就可以让任何两个陌生人认识。”他认为,任何两个陌生人都可以通过“亲友的亲友”建立联系,而两人之间的中介大约是 5人。

在 Milgram之前,麻省理工学院的政治学家索拉·普尔和 IBM的数学家曼弗雷德·科臣曾经做过相关计算,得到的数字是 3。

自称为“实验主义者”的 Milgram对这一数据并不满意,于是亲自设计并执行了著名的六度分隔实验。他从内布拉斯加州和堪萨斯州招募到一批志愿者,随机选择出其中的 300人,请他们邮寄一个信函。信函的最终目标是 Milgram指定的一名住在波士顿的股票经纪人。由于几乎可以肯定信函不会直接寄到目标,Milgram就让志愿者把信函发送给他们认为最有可能与目标建立联系的亲友,并要求每一个转寄信函的人都回发一个信件给 Milgram本人。出人意料的是,有 64封信最终到达了目标股票经济人手中,在发表于 1967年 5月《今日心理学》上的论文中,他描述了一份文件是如何仅用 4天时间就从堪萨斯州的农场主手中转交到麻省坎布里奇某神学院学生妻子手中的:农场主将文件交给一个圣公会教父,教父将其转交给住在坎布里奇市的一位同事,然后文件就到了神学院学生妻子的手中 —— 整个过程只需要 3步,而中间人只有两个。并不是每一个实验对象都如此成功,但平均所需中间人的数目为 5。也就是说,对于 Mil-

gram的陌生人，6步是最远的距离。

虽然 Milgram的实验结果由于设计和操作上的缺陷，受到了一些心理学家的质疑，但是其他实验结果也表明，世界的确不大。在互联网的时代，人们不再习惯通过邮局寄信，可以改用电子邮件重复 Milgram的实验。2002年，美国哥伦比亚大学的研究人员向 166个国家的 6万多网民发去一封连环信，请他们转给随机选中的位于 13个国家的 18名收信者之一，结果发现大部分信件在转了 5～7次后就寄达收信人。2007年，微软研究人员对 2亿 4千万名 MSN用户的 300亿条短信进行分析，发现 MSN用户之间的距离是 6.6步。

世界是如此的小，因为这并不是一个有序的世界。如果世界是有序的，人与人之间的距离有时会非常遥远。如果你要把一个围棋子从棋盘的一端很有秩序地沿着连线一步一步地移到棋盘的另一端，将会有很多步。但是如果在移动时，允许走捷径一步跳到远处的点，就会很快地抵达目的地。在现实世界中，人们的交往有一定的秩序（例如有相似背景的人容易相互认识），组成朋友小圈子，但是也时不时会结识其他朋友圈的人——正是这些“捷径”让世界变得很小。在这个“小世界”中，如果有 3亿人（等于美国人口的 90%，假定剩下的 10%为忽略不计的儿童），每人认识 30个亲友，那么可以算出人与人之间的距离是 5.7。如果有 60亿人（等于世界人口的 90%），人与人之间的距离则是 6.6。

我们的研究表明，在结点数为 4017的汉语句法网络中，结点之间的平均路径长度＜d＞为 3.372，比“六度分隔”还要小，因此，我们认为，语言的句法网络是一个小世界。我们进一步把汉语的句法网络转化为语义网络，发现汉语语义网络的平均路径长度＜d＞为 3.952，聚集系数 C为 0.079，也可以算是一个小世界网络[①]。

这些研究成果分别在 2008年的 Physica A和 Europhys Letter[②]上，引起了国际物理学界的关注。我们还在《科学通报》2009年 7月 54卷 14期上发表了《汉语语义网的统计特性》，介绍了在语义网络方面的研究成果[③]。

这些研究成果显示了语料库的威力，证明了我们确实可以从语料库中挖掘到有用的语言学知识。

语言学知识究竟在哪里？我们的回答是：语言学知识固然在词典里、在语法书里、在汗牛充栋的语言学著作里，但是，这些语言学知识毕竟是通过语言学家对于局部的语言现象归纳出来的，难免会有片面或错误的地方；更多的语言学知识还隐藏在语料库里，语料库是语言学知识最可靠的来源。从语料库中获取语言学知识，并根据这些知识对于前辈语言学家根据内省得出的结论进行检验，从而证实或证伪这些知识，这是生活在 21世纪的语言学家责无旁贷的任务。

除了使用语料库挖掘语言学知识之外，还可以使用语料库挖掘非语言学的知识。

2.从语料库中挖掘非语言学知识

这里我们介绍“文本数据挖掘”（text data mining，简称 TDM）。文本数据挖掘目的在于从大规模真实文本数据中发现或推出新的信息，找出文本数据集合的模型，发现文本数据中所隐含的趋势，从文本数据的噪声中分离出有用的信号。

“文本数据挖掘”（TDM）中的“挖掘”（mining）这个单词并不是一个很确切的比喻。在这里，“挖掘”意味着从没有价值的岩石中提取贵重的金属。如果文本数据挖掘真的遵照这一比喻的话，那就意味着文本数据挖掘是在数据的清单当中寻找新的事实，文本数据挖掘目前倾向于自动地或半自动地发掘大量数据中隐藏的趋势和模式，就像从没有价值的岩石中提取贵金属一样，这样的数据挖掘通常以决策制定为目的。

在十年前，Don Swanson证明了医学文献的语料库中暗含的因果链可以帮助我们找到有关罕见疾病起因的假说，而其中一些假说得到了实验数据的支持。

例如，当调查偏头痛（migraines）的起因时，Don Swanson从生物医学文献的文章标题中提取了各种各样的线索，其中的一些线索如下：

因果链 1：

• Stress is associated with migraines；

（偏头痛与精神紧张有关）

• Stress can lead to loss of magnesium；

（精神紧张可能会导致镁流失）

因果链 2：

• Calcium channel blockers prevent some migraines；

（钙通道阻滞剂可以防止某些偏头疼）

• Magnesium is a natural calcium channel blocker；

（镁是一种天然的钙通道阻滞剂）

因果链 3：

• Spreading cortical depression is implicated in some migraines；

（传播皮层抑郁与某些偏头痛有联系）

• High levels of magnesium inhibit spreading cortical depression；

（高含量的镁可阻止传播皮层抑郁）

因果链 4：

• Migraine patients have high platelet aggregability；

（偏头痛患者有很高的血小板聚集）

• Magnesium can suppress platelet aggregability；

（镁能抑制血小板聚集）

根据这些线索可以假定，缺镁可能是某些偏头痛的原因之一；但是，在 Swanson发现这些链接之前，这一个假定在文献中并不直接存在，它是隐含在文献中的。这个假说还需要进行非文本手段的检验，不过，重要的是，这项研究说明，一个新的、可能是正确的医学假说是来源于文本片段，一旦这个假设得到研究者的医疗专业知识的印证，就可以发现新的医学知识。Don Swanson的研究生动地说明了文本数据挖掘在新知识发现中的重要作用。

我们还可以把文本数据挖掘的技术应用于网络数据挖掘（Web Data Mining）。网络数据挖掘有两个目标。第一个目标是帮助用户在网页上找到有用的信息并在网页文件集描述的范围内找到有用的知识。第二个目标是分析基于网

页系统下的交互，优化系统，并找出用户使用系统的信息。我们实际上是把网页中的信息看成是一个庞大的知识库，我们要从中提取出新的、前所未有的信息。

3. 从语料库中自动抽取信息

从自由文本的语料库中自动地识别特定的实体（entities）、关系（relation）和事件（events）的方法和技术，叫做“信息抽取”（information extraction，简称 IE）。

随着计算机的普及以及互联网（WWW）的迅猛发展，大量的信息以电子文档的形式出现在人们面前，这样的电子文档，实际上就是一个海量的语料库。为了应对信息爆炸带来的严重挑战，迫切需要一些自动化的工具帮助人们在海量的信息源中迅速找到真正需要的信息。信息抽取研究正是在这种背景下产生的。

一般来说，信息抽取系统的处理对象是自然语言文本尤其是非结构化文本或半结构化的文本。但广义上讲，除了电子文本以外，信息抽取系统的处理对象还可以是语音、图像、视频等其他媒体类型的数据。

这里我们主要介绍两种类型的信息抽取：一种是“名称自动抽取”（extraction of names）。一种是“事件自动抽取”（extraction of events）。

语言结构的传统处理方式很少注意名称、地址、数词短语等，我们把它们统称为“名称”（names）。因为这些名称在文本中的分布往往是不均衡的，随着文本的不同而有很大的差异，语言学家认为它们在语言学上没有很大的价值。语言分析中，仅仅是将文本中的单词标注为名词、动词、形容词等，也不注意名称。但事实上，许多文章都包含大量的名称，如果自然语言处理系统不能将它们识别为语言单位，那么就很难对文章做语言分析。不同类型的文章包含不同类别的名称。化学文章中包含化学物品名称，生物学文章中包含与物种、蛋白质及基因有关的名称，报刊中包含大量的人名、机构名及地名。

“名称自动抽取”也就是要对文本中的名称进行自动识别（recognition）和标注（tagging）。

我们将查找人名、机构名和地名作为名称识别和标注的示例。名称识别和分类处理的结果采用标准通用置标语言（Standard Genaralized Mark-up Language 简称 SGML）来标记，在名称开头使用 <NAME TYPE = xx>，结尾使用 </NAME>。

这样，句子“Capt Andrew Ahab was appointed vice president of the Great White Whale Company of Salem-Massachusetts”可以标注如下：

Capt <NAME TYPE = PERSON > Andrew Ahab </NAME> was appointed vice president of the <NAME TYPE = ORGANIZATION > Great White Whale Company </NAME> of <NAME TYPE = LOCATION > Salem </NAME>, <NAME TYPE = LOCATIN > Massachusetts </NAME>

这种分类的基本理念十分简单：我们书写大量的有限状态模式来进行名称的识别和分类，其中每个都记录了名称中的子集并将其分类。这些模式中的内容会根据自身的特性与特定标记分类进行匹配。我们使用标准普通表达符号，特别使用‘+’后缀符来与其中一项元素的一个或多个举例匹配，例如，表达式

Capitalized-word+‘Corp.’

可以表示以大写字母开头并包含一个或多个单词的公司名称。

同样地，表达式

‘Mr.’capitalized-word+

可以与用 Mr.开头的单词序列匹配，并被归类为人名。

要创建一个完整的名称标注器（name tagger），就要编制一个标记文本的程序，然后从文本中的每个单词开始与所有表达式进行匹配；一旦一个匹配成功，单词序列就会被归类，然后再继续此步骤。

如果模式匹配是以特定指向或规则开始的，特别是按照最长匹配或按照给不同规则制定优先权，就必须选择一项最佳匹配。

一个操作性能很高的名称标注器需要一系列的单词列表，例如，一些知名公司名称的列表（IBM，Ford）以及常见首字母列表（Fred，Susan）。

另外，名称标注器还应该具备一个能识别不同别名的装置；例如，在同一篇文章中都出现了‘Fred Smith’和‘Mr. Smith’，这两个名称可能指的是同一个人。‘Robert Smith Park’可能是一个人名或地名（公园的名称），但如果在接下来的句子中出现‘Mr. Park’这样的人名，那么就可以肯定‘Robert Smith Park’也是一个人名。

系统地添加这样的模式和功能，通过机器学习的方法，可以自动训练出一个高效能的名称标注器。然而，这是一个非常艰苦的过程，需要设计一个高水平的系统训练程序。

在对英语新闻的特定话题进行训练和测试时，名称标注器的标注精确度可达到 96%，在对英语新闻的不同话题进行训练时，名称标注器的精确度也达到了 96%。

要知道怎样进行训练，我们来考虑一项简单的任务——人名标注。在人名标注时，每个标记 tag_i 具备 5个可能性：人名的开始、人名的中间、人名的结尾、单个人名的开始和结尾，或非人名。当给一个单词进行标注时，每个单词 w_i 都可能属于这 5个可能性中的一个，为此我们需要计算 w_i 标注为 tag_i 的概率 $p(tag_i | w_i)$。如果 w_i =‘John’，那么，它的 tag_i 就是人名的开始，或者是单个人名的开始和结尾；如果 w_i =‘eat’，那么，以上的两种可能性都为零。对于句子中的每一个单词，都计算它的 $p(tag_i | w_i)$。根据训练的结果，对于一个新的句子，就可以利用搜索算法来求这个句子中可能性最大的人名标记序列，从而从新的句子中抽出人名。

在上面的名称标注中，名称的概率仅取决于当前词，这样的概率是不准确的。在单词‘Mr.’后面可以预测出是一个人的名字，而在单词‘says’的前面也可以预测出是一个人的名字。这意味着，一个标记的概率取决于前面的单词（w_{i-1}）、当前词（w_i）、后面的单词（w_{i+1}），也就是说，我们有必要计算

概率 $P(tag_i | w_{i-1}, w_i, w_{i+1})$，这样，我们就需要使用二元语法（bigram）了。

名称是自然语言中常见的语言单位，大多数的文本都充满着名称，因此，名称的自动抽取就成为自然语言分析的重要的步骤。例如，在事件抽取和机器翻译中，首先都需要进行名称的自动抽取。在基于术语的文档检索中，如果连续的两个单词不是名称，在一般情况下就有必要对它们进行切分，分别处理；而如果连续的两个单词是名称，那么，就应当把它们结合在一起进行处理。在文档标引时，如果把名称分为人名、机构名和地名，索引就可能具有更大的实用价值。由此可见，名称的自动抽取对于自然语言处理具有重要的作用。

在自然语言处理中，名称的自动抽取又叫做“命名实体识别”（Naming Entity Recognition）。一般来说，命名实体识别的任务就是识别出待处理文本中三大类命名实体（实体类、时间类和数字类）和七小类命名实体（人名、机构名、地名、时间、日期、货币和百分比）。其中时间、日期、货币和百分比的构成有比较明显的规律，识别起来相对容易，而人名、地名、机构名的用字灵活，识别的难度很大，因此命名实体识别通常指的是人名、地名和机构名的识别。我们在上面只是介绍了人名的识别，地名和机构名的识别还没有涉及。

命名实体识别的过程通常包括两部分：①实体边界识别；②确定实体类别（人名、地名、机构名等）。英语中的命名实体具有比较明显的形式标志（即实体中的每个词的第一个字母要大写），所以实体边界识别相对容易，重点是确定实体类别。

“事件自动抽取”的主要功能是从文本中抽取出特定的事实信息（factual information）。比如，从新闻报道中抽取出恐怖事件的详细情况：时间、地点、作案者、受害者、袭击目标、使用的武器等；从经济新闻中抽取出公司发布新产品的情况：公司名、产品名、发布时间、产品性能等；从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方，等等。通常，被抽取出来的信息以结构化的形式描述，可以直接存入数据库中，供用户查询以及进一步分析利用。

事件自动抽取系统要从文本中自动地提取某种类型的关系实例或事件。

例如，对于下面的句子：

Harrier Smith, Vice President of Ford Motor Corp., has been appointed President of Daimler Chrysler Toyota.

经过事件抽取之后，我们可以得到如下的两个数据库记录：

记录 1

Person: Harrier Smith
Position: vice president
Company: Ford Motor Corp
Start/leave job: leave job

记录 2

Person: Harrier Smith
Position: vice president
Company: Daimler Chrysler Toyota
Start/leave job: start job

第一个记录是 Harrier Smith在 Ford Motor Corp 离职的记录，第二个记录是 Harrier Smith在 Daimler Chrysler Toyota 就职的记录。

用信息抽取的术语来说，我们从上面的文本中创建了两个填充好的“模板”（templates），而模板中的填充项叫做“槽”（slot）。

我们可以使用正则表达式来描述上面的事件：

capitalized-word+, 'appointed' capitalized-word+, as 'President'

编号 1　　编号 2　　编号 3

与这个正则表达式相应的模板如下：

模板：

Person: 编号 2
Position: 编号 3
Company: 编号 1
Start/leave job: start job

模板中的编号项目可以用与其相匹配的相关编号的文字来填充。

这个模板可以处理如下的简单句子：

Ford appointed Harrier Smith as President

这样的模板还难以处理真正的复杂文本，因为可能出现的句子的变化花样很多，这样简单的模板是难于应付的。这些变化举例如下：

• 公司的名称：Abercrombie and Fitch appointed Harriet Smith as President

• 公司的描述：IBM, the famous computer manufacturer, appointed Harriet Smith as President

• 句子的修饰语：IBM unexpectedly appointed Harriet Smith yesterday as President

• 时态：IBM has/will appointed Harriet Smith as President

• 从句结构：Harriet Smith, who was appointed as President by IBM...

• 动词名物化：IBM announced the appointment of Harriet Smith as President

• 职位的名称：IBM appointed Harriet Smith as Executive Vice President for networking.

• 连词：IBM declared a special dividend and appointed Harriet Smith as President

• 所指照应：IBM has made a major management shuffle. the company appointed Harriet Smith as President this week.

• 必要的推理：Thomas J. Watson resigned as President of IBM, and Harriet Smith succeeded him.

从原则上说，每增加一种变化就需要适当地增加事件模板的“槽”，其结果常常使得模板变得非常复杂。

为了解决这样的复杂性问题，可以使用名称标注器对于文本中的句子进行简单的句法分析，标注时不是使用具体的单词而是使用词组类型符号（如名词词组 noun phrase，动词词组 verb phrase等）来建立模板。例如，对于句子 Ford Motor Company has appointed Harriet Smith, 45, as President

名称标注器可以产生出如下的结构成分：

Ford Motor Company has appointed Harriet Smith, 45, as President

name type = org　　name type = person

通过名词词组（np）分析，可以得到：

Ford Motor Company has appointed Harriet Smith, 45, as

np head = org　　np head = person

president

np head = president

通过动词词组(vp)分析,可以进一步得到:

Ford Motor Company has appointed Harriet Smith, 45, as

np head=org　vp head=appoint　np head=person

President

np head = President

最后,我们就可以得到事件(Event)的描述如下:

Ford Motor Company has appointed Harriet Smith, 45, as President

Event　　person = Harriet Smith　　position = president

company = Ford Motor Company　start/leave job = start job

图 4　事件描述

这样的句法分析结果可以使用自底向上的浅层分析方法来实现。

根据句法分析的结果,我们不难得到如下的模板:

模板:

person = Harriet Smith
position = president
company = Ford Motor Company
start/leave job = start job

上述事件抽取的过程是:

文本→名称识别→np识别→vp识别→事件识别→

图 5　事件抽取过程

通过句法分析得到输入文本的某种结构表示,如完整的分析树或分析树片段集合,是计算机理解自然语言的基础。

作为一种自然语言处理系统,信息抽取系统需要强大知识库的支撑。在不同的信息抽取系统中知识库的结构和内容是不同的,但一般来说,都要有:一部词典(Lexicon),存放通用词汇以及领域词汇的静态属性信息;一个抽取模式库(Extraction Patterns Base),每个模式可以有附加的(语义)操作,模式库通常也划分为通用部分和领域(场景)专用部分;一个基于知识本体(Ontology)的概念层次模型,通常是面向特定领域或场景的,是通用概念层次模型在局部的细化或泛化。除此之外,可能还有篇章分析和推理规则库、模板填充规则库等。

信息抽取系统通常是面向特定应用领域或场景的。这种领域受限性决定了信息抽取系统中用到的主要知识是所谓的浅层知识。这种知识的抽象层次不高,通常只适用于特定应用领域,很难在其他领域复用。如果要把一个信息抽取系统移植到新的领域或场景,开发者必须要为系统重新编制大量的领域知识。

一般说来,手工编制领域知识往往是枯燥的、费时的、易错的,费用较高,并且只有具有专门知识,熟悉系统的设计与实践的人员才能胜任这种工作。

根据"齐夫定律"(Zipf's law),自然语言中普遍存在着"长尾"综合效应(long tail syndrome)。

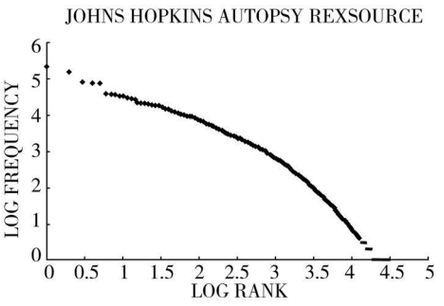

图 6　表示"齐夫定律"的曲线

在表示"齐夫定律"的曲线的后面一段往往拖着一条"长尾"。根据"齐夫定律",自然语言中的绝大多数事实采用经常出现的、非常少量的表达方式来描述,而剩余的事实却需要大量的、不经常出现的表达方式才能覆盖,因此在曲线中出现"长尾"。在这种情况下,人工编制的知识库很难达到很高的语言覆盖面。因此,知识获取问题已经成为制约信息抽取技术广泛应用的一个主要障碍。它除了影响系统的可移植性外,也是影响系统性能的主要因素。正因为如此,近几年召开的多次专题学术研讨会都是以解决知识获取问题、建立具有自适应能力的信息抽取系统为主题的。

目前信息抽取已经成为自然语言处理领域一个重要的分支,通过系统化、大规模地定量评测推动了这项研究不断地向前发展,部分句法分析技术、知识工程研究以及软件工程技术等,都极大地推动了信息抽取研究的发展,迫使信息抽取的研究人员面向实际的应用重新考虑他们的研究重点,开始重视解决以前曾被忽视的一些深层问题,如语义特征标注、共指消解、篇章分析,等等。

目前,系统性能和系统可移植能力是影响信息抽取技术广泛应用的两个最主要的因素。今后信息抽取研究将紧紧围绕如何克服和解决这两个问题展开,重点解决知识获取、篇章分析、高效句法分析等问题,不断提高信息抽取系统的性能,并进一步增强其可移植能力。

4. 语言学研究正在面临战略转移的重要时刻

20世纪 90年代以前,从事计算语言学研究的绝大多数学者,都把自己的目的局限于某个十分狭窄的专业领域之中,他们采用的主流技术是基于规则的句法—语义分析。尽管这些应用系统在某些受限的"子语言"(sub-language)中也曾经获得一定程度的成功,但是,要想进一步扩大这些系统的覆盖面,用它们来处理大规模的真实文本,仍然有很大的困难。因为从自然语言处理系统所需要装备的语言知识来看,其数量之浩大和颗粒度之精细,都是以往的任何系统所远远不及的。而且,随着系统拥有的知识在数量上和程度上发生的巨大变化,系统在如何获取、表示和管理知识等基本问题上,不得不另辟蹊径。这样,就提出了大规模真实文本的自动处理问题。1990年 8月在芬兰赫尔辛基举行的第 13届国际计算语言学会议(即 COLING'90)为会前讲座确定的主题是:"处理大规模真实文本的理论、方法和工具",这说明,实现大规模真实文本的处理将是计算语言学在今后一个

相当长的时期内的战略目标。为了实现战略目标的转移，需要在理论、方法和工具等方面实行重大的革新。1992年6月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议（TMI—92）的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”，就是指以生成语言学为基础的方法，所谓“经验主义”，就是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点。当前语料库的建设和语料库语言学的崛起，正是计算语言学战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法（即经验主义的方法）至少是对基于规则的分析方法（即理性主义的方法）的一个重要补充。因为从“大规模”和“真实”这两个因素来考察，语料库才是最理想的语言知识资源。但是，要想使语料库名符其实地成为自然语言的知识库，就有必要首先对语料库中的语料进行自动标注，使之由“生语料”变成“熟语料”，以便于人们从中提取丰富的语言知识。可以看出，计算语言学现在正在面临着一场战略转移。这场战略转移的关键是知识的获取方式和方法：从依靠“内省”方式转向依靠“语料”的方式，从基于“规则”的方法转向基于统计的方法。

面对计算语言学的战略转移，我们觉得，语言学获取知识的方式方法也应当进行一场战略转移。

与计算语言学相似，传统语言学家获取语言知识的方法基本上是通过“内省”进行，由于自然语言现象充满了例外，治学严谨的学者们提出了“例不十，法不立”（黎锦熙）“例外不十，法不破”（王力）[④]的原则，这样的原则貌似严格，但实际上却是片面的。在成千上万的语言数据中，只是靠十个例子或十个例外就来决定规则的取舍，难道真的能够保证万无一失吗？显然是不能保证的。因此，“例不十，法不立；例外不十，法不破”的原则只是一个貌似严格但实际上是一个很不严格的原则。现在，是抛弃这个原则的时候了。

语料库是客观的、可靠的语言资源，语言学研究应当依靠这样的宝贵资源。语料库中包含着极为宝贵的语言知识，我们应当使用新的方法和工具来获取这些知识。当然，前辈语言学家数千年积累的语言知识（包括词典中的语言知识和语法书中的语言知识）也是宝贵的，但由于这些知识是通过这些语言学家们的“内省”或者“洞察力”发现的，难免带有主观性和片面性，需要我们使用语料库来一一地加以审查。

我们认为，语言学的一切知识，都有必要放到语料库中来检验，决定其是正确的，还是错误的，甚至是荒谬的，从而决定其存在的必要性，决定其是否应该继续存在。我们可以预见，语言学研究战略转移的时代必将到来。一种新的基于语料库的研究方式必将代替传统的依靠“内省”的研究方式，“内省”的研究方式今后只能是基于语料库研究方式的补充，而决不能是语言学研究的主流。从语料库挖掘知识和抽取信息应当是现代语言学的基本研究方法。

计算语言学中的这种战略转移，必将影响到传统语言学。传统语言学也正在面临战略转移的重要时刻。我们应当从高度的历史责任感出发，敏锐地认识到这个战略转移的重要时刻或迟或早总会来临，为此而调整我们的研究方法和研究计划，从而为世界的语言学宝库做出我们中国学者应有的贡献。

英国著名科学哲学家波普尔（Karl R. Popper，1902—1994）在为中文版《波普尔科学哲学选集》[⑤]所撰写的前言中说：“人们尽可以把科学的历史看作发现理论、摒弃错了的理论并以更好的理论取而代之的历史。……我不怀疑我们有许多科学理论是真实的；我所要说的是，我们无法肯定任何一个理论是不是真理，因而我们必须作好准备，有些最为我们偏爱的理论到头来却原来并不真实。既然我们需要真理，……我们除了对理论进行理性批判以外，别无其他选择。”正是本着这样一种对于传统的语言学研究结论进行理性批判的科学精神，我们需要在语料库大量语言事实的基础上进行理性的审视，这样，我们就有可能提出不同的、但更富于发展前景的学术观点。

在语言学研究中，我们尽最大的努力避免偏颇和错误。波普尔在他的同一篇前言中还说：“科学是可以犯错误的，因为我们都是人，而人是会犯错误的。因而错误是可以得到原谅的。只有不去尽最大的努力避免错误，才是不可原谅的。但即使犯可以避免的错误，也是可以原谅的。”语料库给我们提供了极其丰富的语言客观事实，我们应当充分利用语料库给我们提供的语言客观事实，避免前辈语言学根据“内省”的研究方法做出的可能有片面性的结论，从而推动语言学的发展。

与此同时，我们还应当使用文本数据挖掘的技术和信息自动抽取的技术，从语料库中挖掘和抽取非语言学的知识，把语料库的应用扩大到其他的领域，进一步推动自然语言处理研究的发展。

注释：

① Liu H. The complexity of Chinese dependency syntactic networks [J]. Physica A, 2008(387).

② Liu H., Hu F. What role does syntax play in a language network? [J]. Europhys Lett, 2008(83).

③ 刘海涛. 汉语语义网的统计特性 [J]. 科学通报, 2009, 54(14).

④ 王力在《汉语史稿》（上册）（中华书局，1980）中指出，“所谓区别一般与特殊，那是辩证法的原理之一。在这里我们指的是黎锦熙先生所谓‘例不十，不立法’。我们还要补充一句，就是‘例外不十，法不破’。”

⑤ 波普尔. 波普尔科学哲学选集（纪树立编译）[M]. 北京：生活·读书·新知三联书店，1987.

Abstract: This paper describes the discoveries of the complex network of the Chinese language by the doctoral students at Communication University of China. Their work was based on the Chinese dependency Treebank. Also, it introduces cases of nonlinguistic knowledge mining and automatic information extraction by scholars abroad. The writer holds that linguistic studies should be based on knowledge mining and automatic information extraction.

Key Words: corpus; dependency Treebank; knowledge mining; information extraction