汉语语料库的计算机环境

薛业涛
(中国科技大学计算机系，合肥　230027)

一、引言

语料库是一种具有大规模的真实的语言资料的计算机资源。随着语言研究的现代化和大规模真实文本处理的需要，语料库研究越来越引起各方面的重视并得到迅速发展，基于语料库的研究方法试图通过对大量语料进行分析和统计而获得真实的语言知识，从中归纳出完美的规则集，在此基础上进行语言分析和语言生成。

二、汉语语料库的计算机环境 CCEC 的特点

为了生成真实的大规模的语料库，必须有良好的计算机环境，同时为了有效地利用语料库，必须对语料库的生语料进行加工，生成熟语料，这种加工至少应包括语言元素的切分。对这些元素的进行语言、语法、语义、语用特征的标注，这种加工的工作量浩大。不可能完全由人工完成，应该在一定的环境下，由计算机自动或辅助完成。可见，一个好的计算机环境将使使研究者和用户更好地利用语料库，因此我们设想建立汉语语料库的计算机环境 CECC（Computer Enviroment for Chinese Corpus），具体说，有以下特点：

1.面向语言学研究

语言学的研究是语言工程的基础，其目标是建立适当的语言模型，揭示出详尽的语言规律，这对于建立一个实用的自然语言系统是至关重要的。长期以来，语言学的研究仅限于由语言学家来进行，缺乏各学科的协同合作。语言工作者常常因缺乏合适的语言材料，特别是缺乏大规模的语言文本，而使研究工作局限于某个真实语言的子集中，虽然揭示出一部分语言规律，但是系统缺乏可扩展性，影响了进一步的研究工作。另外，尽管现在越来越多的语言工作者转向基于语料库的研究，但是由于现阶段的语料库只是大量的原始的语言文本，加工的程度很低，使用起来很不方便，妨碍了语言学研究的进展。

为了适应语言研究现代化的要求，为语言工作者提供先进的研究工具和环境，我们建立 CCEC 时，首先要明确面向语言学研究这一基本出发点，原始材料均来源于真实文本，而不是为研究而造出来的句子，同时为了使语言学家能方便地使用语料库，要求：

①为语料库提供查询和统计功能，帮助语言学家从语料库中方便地找到需要的语言现象。

②为了使语言学家无须太多训练就可方便地使用语料库，要求用户环境是可扩充的多窗口图形界面，并用以 Icons、Mouse、Menu 驱动。

③为方便对语料库的增加、删除、修改，语料库的开发环境中必须包括文本处理功能，诸如编辑、合并、编排等，在这种支持环境下，语料库的维护工作就相对容易，并且也可能提供进一步的检索服务。

2.基于知识的语料库

在计算语言学领域，由于受人工智能的影响，长期以来占主导地位的是基于知识的研究方法，这要求计算机掌握足够多的知识，这其中还包括相当多的经验知识及常识，如何表示这些经验知识和常识严重影响一个自然语言处理系统的效果，基于规则的研究方法常常很难获取这些知识，也很难描述清楚这些知识，况且自然语言中有些语言现象本来就是不符合语言理论，这给计算语言学的研究带来了极大的困惑，在这种背景下，我们不得不求助于语料库。如果能建立起一个大规模的语料库，我们就可以认为语料库中包含了足够多的语言知识，我们试图通过语料库去获取这些知识，知识包括两种：一种是事实，另一种是经验，即一种好的猜测和判断。这里我们强调获取经验知识，因为语料库中存储着许多"约定俗成"的语言知识，然后再在这种知识的基础上去检验语法，语义理论，更精确地描述规则。人类知识中最难做到的一点就是如何形成一般性的理性知识，用在语料库中搜索的方法是解决这个问题的一种比较有效的手段，同时还可以根据新的实例，修改过去的理性知识，从一个比较深的层次来看，这实际上是一种基于实例的推理。精确的说，这种语料库环境应该是一种基于知识的语料库，始终以知识为中心。

①在整个环境中，我们仍旧要建立具体的词典和规则库，记录大量来自于真实环境下的语言文本。我们可以认为词和规则库中存放是的理性知识，语料库中存放的是经验知识。用语料库的大量例子去验证词典和规则。分析这些理性知识的正确与否。

②在具体的句法分析中，语料库中应该加入一些最基本的规则，语言工作者经过多年的研究，总结出很多语言规律，其中有相当一部分经过实践检验后证明是正确的，这种规则性知识提示出深层次的语言知识，并且适用面较广，相应的使用频率较高，语料库中使用这些规则后，能大幅度减少语料库的存储消耗。

③语料库不是一个简单的数据库，为了更有效地使用语料库，必须对语料库中的原始语料进行多层次的加工，不断丰富语料库中的内容，这种加工包括自动分词、语音、词法、语法、语义标注。

④现有的机器翻译系统普遍存在着脆弱性，具体来说，当系统的外部环境稍微发生变化时，系统就完全不知所措。一个似然的解决方法是加入足够多的知识，包括常识，以扩大系统的适应能力。但是，无论怎样扩大其知识库，总会有一些意想不到的情况出现。一个彻底的解决方法是让系统自动学习知识或技能以适应环境的变化，即具备不断增长的智能。借鉴机器学习的原理及技术，试图让机器在语料库中自动寻长新的语言知识，这是数据库技术与机器学习相结合的范例。