中文信息处理与汉语研究

詹卫东

中文信息处理（Chinese Information Processing）指的是利用电子计算机来存储，加工和传播以汉字形式承载的信息。中文信息处理大致上可以区分为两个层次：（1）在**符号层**进行的中文信息处理；（2）在**内容层**进行的中文信息处理。

符号层的处理要研究的是汉字在计算机中如何输入、输出；内容层的处理则要研究如何让计算机能像人一样“看懂”中文的意思。举个简单的例子，对于“聪明”这两个汉字，就像人们可以自如地写在黑板上或者纸上一样，现在人们也可以通过键盘输入到计算机中，显示在计算机的屏幕上，这就是计算机在符号层次上进行中文信息处理。在这个层次上，计算机处理的是“信号”，也就是“符号”的“形式”，而不是“内容”。如果可以编写出一个计算机程序，当人们向计算机输入“张三很聪明”时，这个计算机程序可以响应（输出）：“张三脑子蛮灵光的”或者 “Zhang is very smart”，人们就会觉得计算机看懂了“张三很聪明”这句话的意思。这时候，计算机就是在内容层次上进行中文信息处理了。在这个层次上，计算机处理的是真正意义上的“信息”，也就是“符号”的“内容”。

国内真正开始比较系统的中文信息处理研究工作是从 20 世纪 70 年代中期开始的，到现在已经有近 30 年时间了。可以说目前中文信息处理在符号层的处理成果已经完全进入实用，在我们每天的日常工作和生活中，都要使用计算机或其他一些电子设备来输入、输出及传播以数字形式编码的汉字信息。但内容层的中文信息处理还远远没有达到实用阶段，尽管也有一些成果已经进入信息产业市场（比如一些自动翻译软件），但这方面的许多研究工作目前还是实验室中研究人员的科研课题。

实际上，无论是在符号层，还是在内容层，要让计算机能够处理中文信息，都有许多困难需要克服。

先说符号层的处理。跟拼音文字（比如我们熟悉的英文）不同，汉字是大字符集文字体系，从历代字典收字数量可以知道，汉字的基础字数就是非常庞大的，而且还一直在增加，东汉的《说文解字》收录 9353 字，清代的《康熙字典》收录 4 万 7 千多字，当代的《中华字海》更是收录了 8 万 6 千多字。相比之下，英文在计算机中的标准语言——美国信息交换标准编码（ASCII 码）——只有 128 个字符，算上扩展的 ASCII 码也只有 256 个字符，要区分这 256 个字符，在计算机中只需要 8 位二进制数（2^8=256）就可以了。而如果要让全部汉字都能进入到计算机中，用 16 位二进制数的编码空间来为每个汉字进行编码也不够用（2^{16}=65536），尽管这在技术上并不存在障碍，但落实起来，由于软件兼容性，存储空间利用率，汉字排序等等问题，还是会有很多麻烦。此外，“全部汉字”也只是个理论概念，实际上没有人说得清楚“全部汉字”有多少。除了中国大陆使用的简体汉字，还有港台地区使用的繁体汉字，另外还有其他国家，比如日本和韩国使用的汉字。把所有这些汉字都集中起来，纳入到一个统一的汉字编码体系中，显然不是件容易的事情，尽管现在已经有不少汉字编码体系在实际应用中发挥巨大作用，比如中国大陆地区通行的 GB 码，GBK 码，港台通行的 Big5 等，但由于一直以来都还没有一个完全统一的汉字全字符集编码体系，因此汉字编码标准化机构目前仍然还在向信息化时代的“书同文”目标在努力着。除了汉字的机内编码问题，汉字在计算机上的输入方式，字形显示等等方面，目前虽然已基本满足了人们应用计算机的需要，但也还有不少问题有待进一步改进，比如随着信息时代数字化步伐的加快，越来越多的个人电子设备需要用到汉字的输入输出，比如我们日常使用的手机、PDA 等，都碰到汉字输入和显示的问题。这些设备不同于标准计算机系统，往往有存储容量、键盘大小等各种限制，从而也对符号层的汉字处理提出更高的要求。如何高效便利地在更多的电子产品中使用汉字，仍然是在符号层进行中文信息处理需要进一步探索的课题。

再说内容层的处理。在这个层次上考虑中文信息处理问题，也就是从“字”的层次进入到

“语”的层次。要让计算机能够处理汉字所承载的汉语信息，就要求计算机能够理解汉语的词、句子、篇章的含义，这就要求汉语研究工作者能够把对汉语的研究成果尽可能地转化为计算机可以懂得的各种语言知识库，只有配备了相应的语言知识库，计算机才有可能模拟人的语言行为，理解人所说的话。而要做到这些，难度要远远大于在符号层上进行中文信息处理所碰到的困难。归结起来，在内容层上进行中文信息处理，碰到的全部困难都可以概括为语言的各个层次上所存在的歧义问题。跟有形态变化的语言相比，汉语因为缺乏系统的形态变化特征，歧义问题显得更加突出一些。举个简单的例子。汉语的“老师”既可以指一位老师（单数），也可以指多位老师（复数），而英语“teacher”一般是单数，指一位老师，如果是复数，则要以“teachers”形式出现。这就从形态上区分了一些意义，从而为计算机理解词语的意义提供了一些有用的线索。不过，有没有形态特征来为理解语言的意义提示线索还不是最重要的。自然语言（当然也包括汉语）在从词到篇章的各个层级上都存在着许多固有的歧义，有的歧义是人们比较容易察觉的，还有大量的歧义可能是一般人在交流时忽视了的，这种种的歧义问题，在计算机进行自然语言的意义理解时，都会迎面碰到，构成极大的困难。这里我们不妨举几个简单的例子来略做说明。

相信不少人都听过这样一个笑话。老师让学生用“难过”造句，结果学生造出一个这样的句子“我家门前的小河很难过”。这个笑话实际上就是利用语言中的歧义现象造成的幽默效果。“难过”既可以作为一个整体，理解为“难受、伤心”的意思，这个意义上“难过”是一个形容词；也可以分开来理解，其中“难”表示“困难、不容易”的意思，“过”表示“跨过去”的意思，做这种理解时“难过”是一个动词性词组（后面还可以带宾语成分，比如“很难过这一关”）。老师让学生造句，本意是将“难过”作为前一种意思理解时造一个句子，但学生的答案中，“难过”却需要作为后一种意思来理解，从而引起冲突，出了笑话。像这样的歧义人们是比较容易察觉的。还有大量的歧义蕴含在我们的话语中，但却不大为人所知，比如：“这些大学生不看重大城市户口”，相信一般人看了不会觉得会有歧解，但实际上如果把这个句子中分解为若干个词，至少有两种分解方式，一是{这些 大学生 不 看 重大 城市 户口}，还有一种是{这些 大学生 不 看重 大 城市 户口}。一般人们都会正确地按后一种方式来分解，但为什么不能按照前一种方式来分解呢？换句话说，按照前一种方式来分解，所得到的词汇也都是汉语中完全正常的词汇，但为什么这些词汇组织在一起形成的句子就不能理解了呢？当人读报纸或者听别人说话时看到或者听到上面这句话，会“很自然地”理解这句话的意思（这其中就包含了很自然地将这句话分解为若干个词的过程），就好像我们每时每刻都在呼吸空气一样地自然，同样，就像人们很少会停下来思索一下呼吸是一个怎样的过程，人们也很少会去思考——“我们为什么这么自然地就理解了一句话的意思呢”。

但是，在信息时代，计算机开始向人们提出这样的问题了。

现在还没有人能回答得了这个问题，不过，包括计算机科学家，数学家，语言学家在内的许多学科领域的专家学者都开始起来响应信息时代所提出的这个大问题，从不同学科的各个角度来试图发现人自己是如何进行语言信息处理的。只有先搞清楚人自身是如何进行语言交际，如何理解别人的话语，才有可能在计算机上模拟人的语言行为能力，实现真正意义上的，在内容层上进行的语言信息处理。在这个问题上，英语如此，汉语亦然，各国的信息处理专家现在都在朝这个目标积极努力。而在这个过程中，对语言本身结构和意义的研究，比如对汉语词语如何组织成句子，理解句子意义的机制是什么，句子如何组织成篇章的研究等等，将起到举足轻重的作用。传统的汉语研究是面向人的，它更多地是关注如何定性地描写一种语言现象，而很少关注如何去说明人们理解句子意思的过程。现在，面向信息处理来开展汉语研究工作，对传统的语言学研究提出了更高要求和许多新的挑战，同时这也成为摆在每一个语言文字工作者面前的一项紧迫而又有很强时代感的任务。可以毫不夸张地说，未来的中文信息处理的发展将有赖于汉语研究的进展，未来的中文信息处理专家队伍也必然会有更多从事汉语研究的人加入进来。

古老的汉字文明已经在信息时代跨过了数字化的第一道障碍，从纸和笔的世界跻身光与电的行列，现在，信息时代高速前进的步伐在呼唤中文信息化的第二次飞跃——让计算机看懂中

文。在这一征程中，从事汉语研究的语言文字工作者责无旁贷，应该肩负起自己的历史使命。

作者单位：北京大学中文系，北京大学汉语语言学研究中心，北京，100871
电子邮件：zwd@pku.edu.cn　　个人主页：http://ccl.pku.edu.cn/doubtfire/