加强中文信息处理的基础建设

宋柔
北京语言大学

中文信息处理已经走过 30 年，取得很大成就，但并未取得主流学科的地位，根本原因在于：汉字输入输出基本解决以后，

中文信息处理未能持续地对科学进步和社会发展做出重大贡献。

深层次的原因

- 学科本身难度太大
- 基础建设不够

因此，不必怨天尤人，还是要加强主观努力，特别是踏踏实实做好基础建设工作。

1. 加强基础研究

计算语言学是数学、计算机科学、语言学的交叉学科。要想推进计算语言学的研究和应用，提高中文信息处理的社会贡献度，不能仅仅在交叉接口上做文章，必须在应用导向的前提下，深入这三个学科进行研究，在基础研究方面取得突破。

（1）数学—模型

- 各种统计模型的数学性质（适用范围的约束条件，这种条件与实际文本的吻合程度及对应用效果的影响，比如某种特征发生的随机性和独立性）
- 建立更加适合中文信息处理的数学模型
- 语料库的数学性质（语料库规模、语料库质量、特征规模和深度、语料库建设成本等因素的定量关系，语料领域差异和文体差异的定量研究）
- 容错统计（语料库的文本字符错误、词语切分错误、标注错误，模型与实际不符，导致统计数据可靠性不高，应用效果受影响，需对这种影响作出定量估计以便制定应对策略）

（2）计算机科学—算法

- 巨大规模生语料的统计，特别是多元字符串组按照与或非组合的逻辑关系隔距共现的算法
- 各种数学模型的高效率的实现算法

（3）语言学—形式化机械化的语言研究（同语言学界的研究目标、研究方法很不相同）

- 字（例如：在字符集开放并涉及非标准书写的条件下，汉字字形形式化描述体系和比对算法）
- 词语（例如：词语的界定，词语属性和分类，同形异质词语的界定和区分）
- 句和篇章（例如：句和超句的界定，句法句义分析，句间关系分析，篇章的话题结构和逻辑结构）

2. 加强基础数据建设

（1）建立国家级的词语库，包括

- 通用词语库
- 领域术语库
- 地名、人名、企业名、商品名库

不要拘泥于词或分词单位的概念，凡有独立意义的 2 字组、3 字组、4 字组都收，再加

上难以用短语规则分析的更长字组，从而将分词中未登录词识别问题转化为切分歧义处理问题，。

每个词语条目至少应包括字符串形式、读音、简单解释、例句，最好包括同形异质词语的列举、解释，还最好有多语翻译（包括少数民族语言）。

采用互联网合作方式建设。

（2）争取开放语言资源（包括各种文本、词典）供研究使用

各种书刊报纸（包括双语对照书籍、各种词典）的文本的数量巨大，利用价值极高，但受制于知识产权保护而闲置，非常可惜。此事正是发挥社会主义制度优越性的最佳范例。

需争取知识产权界的认可和支持。

3. 加强管理体制的基础建设

目前的主要障是成果和人才水平的评价体系过于数量化，逼迫研究人员急功近利，对于交叉学科影响更为严重。中文信息处理不是一般的交叉学科，而是文科理科大交叉，受害更大。比如，研究成果难以在高等级主流专业刊物上发表，ACL 等国际会议文章不被计算机学界认可，语言学刊物的论文更难以纳入自然科学基金课题结题成果范畴，等等。此事需要学会领导做工作，取得学术界的共识，争取科研和教育管理部门对于文理大交叉学科发展的支持。

4. 学术带头人必须亲自做第一线科研工作

中文信息处理，如同其它具有足够难度的工作一样，没有巨大的付出就不可能取得巨大成就。30 年的中文信息处理实践证明，50 年的自然语言处理实践也同样证明，没有任何窍门，没有任何巧妙技术和模型，能让我们绕开其核心困难而取得根本性的突破。要想对核心困难有所破解，必须有一批具有相当功力的人去攻关（并非公关！）。

六零后、五零后的科研将帅，必须有足够多的时间从事第一线的科研工作：

- 自己收集收据。
- 自己整理数据。
- 自己标注数据。
- 自己设计程序实现自己提出或自己使用的某个模型。
- 自己分析实验结果，去粗取精，去伪存真。
- 自己提出改进方案并去实现。

研究生的实验数据往往是靠不住的。

坐在第二线、第三线是号不准脉的。