国家自然科学基金项目“规则与统计相结合的现代汉语虚词用法自动识别研究（60970083）”
国家自然科学基金项目“隐喻识别与理解的理论与方法研究（61170163）”
蒋经国国际学术交流基金“历代语言知识库建置”计划（2009）

中文信息处理战略研讨会2012

深度计算与语义计算

俞士汶　朱学锋
北京大学计算语言学教育部重点实验室
北京大学计算语言学研究所
Email: yusw@pku.edu.cn

2012年4月14日，江西婺源

主要内容

- 关于研究思路的源头
- 中文信息处理的最高境界
- 多模态信息的融合
- 互联网环境与深度计算
- 语义计算的发展脉络
- 结语与致谢

关于研究思路的源头

研究题目是怎么想出来的？动力来自何方？

（1）社会的需求：政府部门科研计划指南，实业单位的需求——功利性

（2）学科的终极目标——科学追求 / 兴趣

（3）原有基础上的更上一层楼——现实性

（4）同行启发（论著、讲演、交流、评测）——群体性

多种因素共同作用，中文信息处理研究得以呈现繁花似锦的灿烂景象。

中文信息处理的最高境界

自然语言理解可望而不可及？

自然语言处理是数值计算机在非数值领域最早的应用（MT，Turing试验），一直未能取得突破性进展。困难所在：

（1）人类对自身的语言机制不甚了了；

（2）语言既是对象，又是工具；

（3）无限的语言现象与有限的计算资源。

自然语言理解能走多远？ 如何走？

多模态信息的融合

目前的计算语言学和自然语言处理的研究只着眼于话语或文本，无论是采用规则方法还是统计方法。但是，实际上人类不是孤立地在使用语言。人与人的交际是多通道的，利用多模态信息，包括对实物的认识和感知、语音、图形、影像、记忆与知识等等，文字只是一种形态。人的脑、口、眼、耳、身体并用，实现了多模态信息的融合与互助。部分功能障碍者（盲、聋、哑）就会遭遇各种不同的困难。

关于人类智能本质研究的**A，B，C**

（A：Artificial intelligence，B：Brain science，C：Cognitive science）

中与**B，C**（神经语言学、认知语言学）相关的学科已在关注、探索这些问题。属于**A**的自然语言处理似乎还没有太关注这些问题。原因可能是计算机的能耐太大了，也可能是实用的需求过于强烈了。可以归属于 **A**的最新成就,如**2011**年**IBM**的 **Watson**系统，**Apple**的**iPhone4S(Siri)** 以及**2012**年**Google**的数字眼镜。这些进展能否说自然语言理解取得了突破性进展？是否实现了多模态信息的融合?

多模态信息的融合 一些零星的想法和见到的报道

对大脑与语言的机制的认识

脑常常被视作进化过程中的最高成就，因为它赋予了人类一些高级特征，例如 语言、智慧、意识。语言是进化的终极发明。在令人类区别于动物的特征中，语言处于核心地位。语言也许称得上是人类的决定性特征之一。（英国《新科学家》周刊 2005年4月9日 的文章）

语言系统是动物进化到人的两大标志之一。

长期以来，虽然认为汉语信息处理技术有其特殊困难，从理解高度看，各种语言应该是共性远超个性。但最近看到不同的报道。

婴儿拥有学习任何语言的普遍潜能，但他们最终使用的语言对大脑神经传导途径的形成产生重要影响。神经系统语言回路最终的组成方式完全取决于它们所支持的语言类型。从小到大说中文的人的大脑肯定与只说英语的人的大脑不同。（香港《南华早报》网站2011年10月9日文章）

如何模拟大脑和语言的机制?

大脑马里兰大学正在研究一种脑机接口技术，一种非侵入性的、内衬传感器的“脑帽”。通过神经接口软件，可利用脑电图(EEG)阅读脑电波，并将脑电波翻译成计算机和其他设备可执行的指令，如控制计算机、假肢、轮椅甚至数字化的“阿凡达”。（美国每日科学网站2011年7月29日报道）

多模态信息的融合 一些零星的想法和见到的报道

“读心”设备可能成为现实。一种可以读出大脑受损患者想法的设备有望成为现实。科学家们已证实，电脑能够通过分析患者脑电波“猜测”到患者听到的是什么单词。大脑把听到的词分解成复杂的脑电活动模式，后者能被解码并翻译成与原来听到的声音相似的版本。大脑处理思维的方式与处理声音的方式相似，因此他们希望这一突破性发展能帮助他们研发出一种植入物，可以诠释无法说话的患者脑子里想说的话。

（英国《每日电讯报》2012年1月31日报道）

一个国际联合科研小组在西班牙马德里理工大学启动了一项“人类大脑计划”，旨在通过超级计算机模拟人类大脑灰质的运转，了解人类神经元之间如何相互联系。科学家称这项研究堪比登月。人类对大脑的研究200多年前就开始了，论文数以百万计，人类对大脑的认识仍很肤浅。

（西班牙《阿贝赛报》2100年5月10日报道）

神经操控将成为战争利器.基于人脑的技术多种多样。不久，“人脑-机器接口”技术可能会把人脑与计算机程序结合起来。

（美国《科学新闻》周刊网站2011年11月11日文章）

让意识与躯体脱离，让人类实现“永生”。俄大亨欲把人类思维移植给机器人。第一阶段——人脑控制的机器人——已经指日可待。

（美国《科学大众》月刊网站2012年3月4日文章）

多模态信息的融合

973项目“数字内容理解的理论与方法”曾试图探索这个问题。

顾曰国教授建立了记录实际场景的现场即席话语多模态语料库（包括话语活动的音频、视频文本及其转写的文字）。

手语机器翻译研究的启示（2011年5月11日，ICL/PKU手语研究讲座）。

南京师范大学有一个语言科技研究所，把相关研究实体集合在一起，已有语音设备、脑电波仪器、眼动议等设备（与医院合作利用核磁共振设备）。

我的认识：自然语言理解必须仰仗脑科学、认知科学的进步，多学科的交叉和融合才有希望。我们应该发出一些声音，做一些努力。语言学（计算语言学）也会有自己的贡献，特别是语义计算研究是向自然语言理解进军途中的一支重要的方面军。

互联网环境与深度计算

今年973项目指南中有“互联网环境的中文深度计算”这样一个题目。我觉得出得很好。

计算语言学发展环境的变迁

有限词汇与典型句型——〉大规模真实文本——〉互联网

语言深度计算与互联网的关系: 面向乎?基于乎?

从实用的角度看，可能更关注“面向”。从研究的角度看，应该更关注“基于”。

关于深度计算的含义:既是“深度”，就包含了“由浅入深”，继承已有研究与成果。“深度”不是某个“刻度”，覆盖一定的范围，且可不断深入。当前会聚焦于语义计算。

语义计算的发展脉络

——从不同角度考察

按作为计算对象的语言单位划分，由小到大：

词语（概念/词汇语义学）

↓

语句（框架/句法语义学）

↓

篇章（情境/篇章语义学）

按对内容理解的程度，可分互有联系、相互支持的3个层次：

本体层次上的语义计算

↓

认知层次上的语义计算

↓

语用层次上的语义计算

本体（ontology）层次上的语义计算

基于知识库（名词的概念层级和动词形容词的语义角色），
本质是借助客观的世界知识（常识）消解语言单位和语言结构的歧义。

各种语言的语义计算的主攻方向。英语领导潮流，以英语为背景，
创立了各种理论、算法。汉语也有一定的成果和积累。

与国际先进水平的差距在缩小？

国家自然科学基金，973，863等国家项目在这方面都提出了新的研究计划，除前面提到的973指南外，还有863已立项的“大规模中文语义信息处理技术与系统”。

可实现互联网上的信息服务向知识服务的提升。

认知层次上的语义计算

——以隐喻计算为例

尽管自然语言理解研究的主攻方向一直是语义歧义消解，但是仅仅消解了歧义，还不能完全解决文本内容理解的难题。一些文学表现手法，像隐喻、影射、双关、夸张、拟人以及遣词造句的技巧对自然语言处理研究提出了挑战——超出歧义范围。甚至，消歧也并非是语言理解的必要任务。

双关的实例：　　**“您的健康是天大的事——天大药业”**

“您　的　健康　是　天　大　的　事”

“您　的　健康　是　天大　的　事”

“一面之缘，终生难忘”

这些使用技巧并非只见于于文学作品，人们日常语言中也经常使用，反映了人类的认知思维机制。

重点讨论隐喻。

认知层次上的语义处理

——以隐喻计算为例

对隐喻（metaphor）的基本认识

各个语言层级上都有隐喻存在：

构词层级：卵石[1]　杏仁眼[0]　人流[1]　美女蛇[0]

词汇层级：潮流[2]　朝阳[2]　燃烧[2]　纯净[2]　蓬首垢面　同舟共济

短语层级：知识[1]的海洋[1] /　播种[1]幸福[1]的种子[1] /金融[1]海啸[1]

句子层级：汽车喝汽油 / 老公是鸡肋[2]

篇章层级：打起黄莺儿，莫叫枝上啼。啼时惊妾梦，不得到辽西。

隐喻不仅具有修辞功能，而且是语言发展和变化的一种重要方式，更是人类一种基本的认知方式，总是借助身边的、熟悉的事物用比较方法认识新事物和形象化地表述事物的重要手段，思维与语言中无所不在。既然运用语言离不开隐喻，自然语言理解研究就必须攻克隐喻计算这个堡垒。

认知层次上的语义处理

——以隐喻计算为例

隐喻计算研究的任务：

（1）隐喻识别

知识的海洋 —— 海洋资源考察

（2）隐喻理解(与翻译)

知识的海洋 —— 知识像海洋一样丰富

老公是鸡肋 —— 老公像鸡肋一样食之无味弃之可惜

（3）隐喻生成

“信息的海洋”，“旗帜的海洋”

目前研究重点放在语句层级隐喻的识别和理解上。

与搜索关系最为密切、也最便于应用的是短语隐喻的识别。

认知层次上的语义处理

——以隐喻计算为例

已经做的和正在做的研究工作

（1）2002年提出研究设想，2004年列为973课题“文本内容理解的数据基础”（2004年9月—2009年12月）的子任务之一。

（2）2006年王治敏完成博士论文《汉语名词短语隐喻识别研究》，已由北京语言大学出版社正式出版。

（3）2007年南京师范大学曲维光博士申请到国家自然科学基金“汉语隐喻理解关键技术研究” （2008年—2010年 ）

（4）2008曲维光博士《现代汉语词语级歧义自动消解研究》出版，其中第10章 “隐喻识别研究”

（4）2009年8月俞士汶应邀在百度技术创新大会上作了特邀报告《隐喻与词义的计算研究及其在搜索引擎中的潜在应用》

（5）2010年5月贾玉祥完成博士论文《汉语文本中的隐喻计算研究》。

（6）2011年王治敏博士申请到国家自然科学基金“隐喻识别与理解的理论与方法研究” (2012-2015)——构建隐喻知识库

语用层次上的语义计算

（1）构式的凸现意义

这一锅饭够吃五个人

这一张床可以睡三个人

台上坐着主席团

语言构式凸现的意义并不等同于成分（中心词）的默认意义。这些构式凸现的是实体与实体之间的数量分配关系、空间位置关系，主要动词与名词间原有的施受关系等虽然存在，但退居次要地位。

（2）语义指向

例如：述补结构、状中结构中的补语、状语的语义指向

（文章）写完了 /（老师）写累了 /（毛笔）写秃了

香喷喷地炸了一盘花生米 / 园园地围成一圈

原有的知识库中的知识不够用，要反映语义角色的变化过程与结果。

可能的解决之道：建设基于广义配价理论的语义知识库——

（1）顾及各类构式的语义关系。

（2）不仅描述动作的参与者，还要描述动作参与者的变化。

语用层次上的语义处理

——源自陆俭明教授的报告

（3）语义和谐律

词语之间语义制约的原则，本质上就是要求句子中的各个词语之间在语义上要和谐。能否说，语言中就存在着“语义和谐律”（semantic harmony ）？

拔出来/ *拔进去/ 插进去/ *插出来

说话和气点儿/ *说话粗暴点儿/ 说话严肃点儿

那个大苹果他都吃了 / *那颗小樱桃他都吃了 / 那颗小樱桃松鼠都吃了

“强壮的钙，聪明的锌。”

相关研究有益于病句剖析和语言自动生成。

原意用“语义计算”而不用“语义分析”的理由。

立足于汉语本体研究成果，特色鲜明， 解决这些问题，更有创新性，而且会对认识语言的本质和共性做出贡献。

结语与致谢

衷心感谢会议给我们发言的机会。浪费了大家宝贵的时间，还请原谅。

座右铭：“路漫漫其修远兮，吾将上下而求索”。

寄希望于同行者和年轻的、更年轻的一代又一代。

谢谢大家。