

# 中文信息处理功能与语言教学

徐波

# 建议主要关键词

## • 学术方向

- 语用理论、汉语本体、语义分析、知识元、篇章分析与理解、中文理解、跨语句理解、认知机理、知识处理、应用语义等
- 少数民族语言、跨语言、民族语言翻译等
- 分词及分词标准（连说就连写）、感情分类、情感分析、智能接口（包括文字识别）、口语处理、汉语句典、句模（**pattern**）等
- 研究范式双语研究、半监督学习机器学习

## • 资源平台

- Web语料、资源库建设、资源共享、平台标准、考虑平台、标准的统一和资源整和、中文信息处理资源仓库、少数民族语言信息化应用层次、标准化、中文资源共享平台、中文标准化工程等

## • 需求驱动

- 知识服务与知识社会，知识检索、文本挖掘等；
- 互联网信息含粗粒、中文教学现代化、中文远程教育平台、家庭数字化工程和中文电脑等

- 语义的问题---路线图如何？
  - 非常困难，需漫长的过程和大规模验证；
  - 困难在于：语义与语用（语境）相关；
  - 工程界：面向领域的知识工程；
  - 需要建立语义与知识工程之间的桥梁；
  - 方有可能建立一个全社会都能参与的语义知识库研制参与机制

# 焦点问题---包括少数民族语言在内的 多语言处理、中文资源共享、标准化

- 核心是资源的共享问题和应用问题；
- ChineseLDC的尝试
- 技术路线：通用+专用+共享
- 通用要有代表性；专用要有针对性；尤其是专用资源库
  - 必须在足够技术研究的基础上
  - 能明确讲清楚建设什么样的专用库后，对领域会产生什么样的影响？

# 焦点问题---知识服务与中文教学

- 从信息技术角度：**Enable**技术，使能技术；或这**GPT**技术，通用技术
- 从语言基本功能看
  - 信息处理---知识服务
  - 沟通交流（人人，人机，人物）--语言障碍

# 沟通交流---需要场景

- 人类口语交流的问题；
  - Computer-mediated Communication between Human (including translation)
- 人类语言的学习问题
  - Computer-aided Language Learning
- 人机/人物对话问题
  - 如口语对话系统
- 最大的问题，如何描述场景或者领域
  - 所谓的domain specific非常困难

# Immersive语言学习和交流

## 语言学习和交流的场景构造



## Immersive environment

Personalization

Active participation

Immersion

## Interactive platform

Dialogue Management

Facial Express

Gesture

.....

## Language Evaluation

Pronunciation

Text and Paragraph

Communication

Machine Translation

Information Retrieval

# 中长期目标

- 信息处理必然进入多语言处理与语义理解阶段
  - 语义方面的重大挑战，需要进行基于场景和领域的知识工程的积累
  - 需要基于现有语义描述体系的规模化验证
- 交流沟通尤其是学习必然进入沉浸式场景沟通阶段
  - 需要综合集成语音语言翻译、处理与评估，与基于场景构造的虚拟现实技术进行结合，营造语言无法形式化描述的场景；
  - Interactive 学习的问题

CASIA

谢谢！

