计算语言学基础

第一讲 引言

王小捷

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

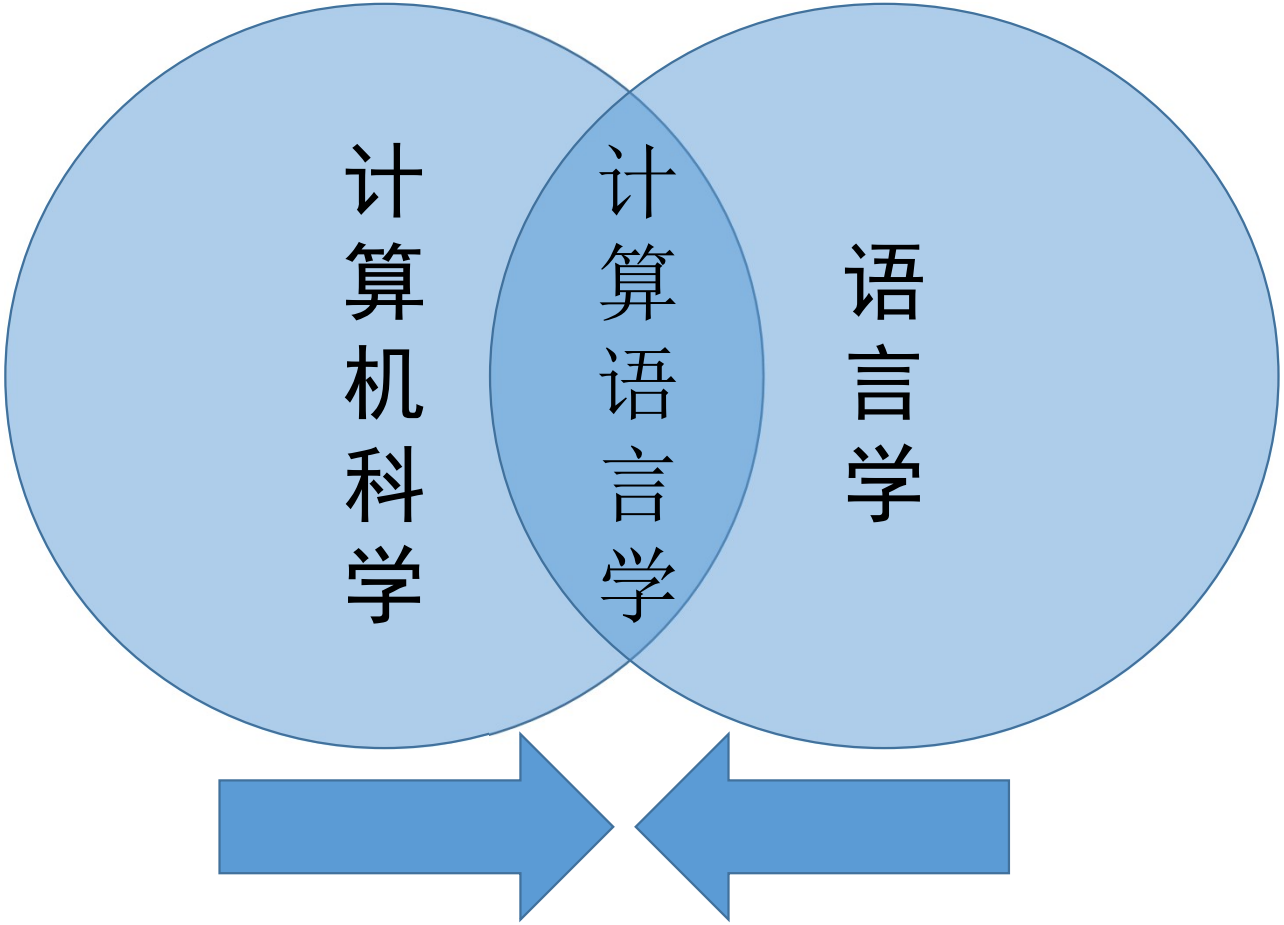
计算语言学(Computational Linguistics)

- 计算 / 语言学
- 从计算视角开展的关于语言的科学研究与应用
  - 研究对象：自然语言
    - 人类自然语言：汉语、英语、手语？盲文？
    - 动物自然语言：？
    - 人工语言：计算机语言

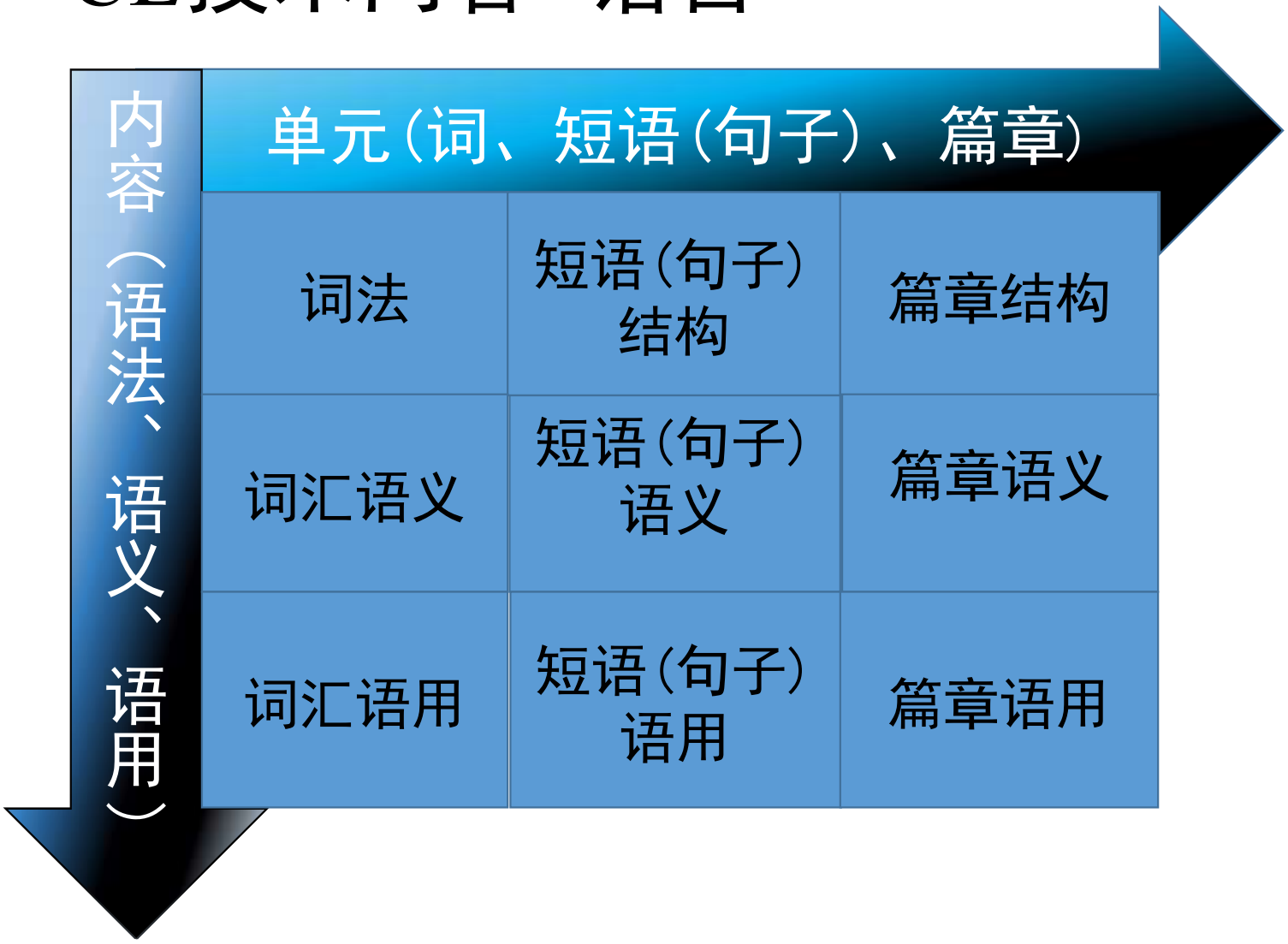
计算语言学(Computational Linguistics)

- 研究目标：为自然语言对象建立计算模型，从而使计算机器具有处理人类语言的能力
  - 语言对象的计算模型
    - were➔are+过去时，计算机器➔计算/机器
    - 嵇康是被司马昭处死的，他为人耿直…
    - ……
  - 计算机器具有处理人类语言的能力
    - 处理的结果像人　vs　处理的过程像人(强)
    - 人机对话(图灵测试)、机器翻译、……

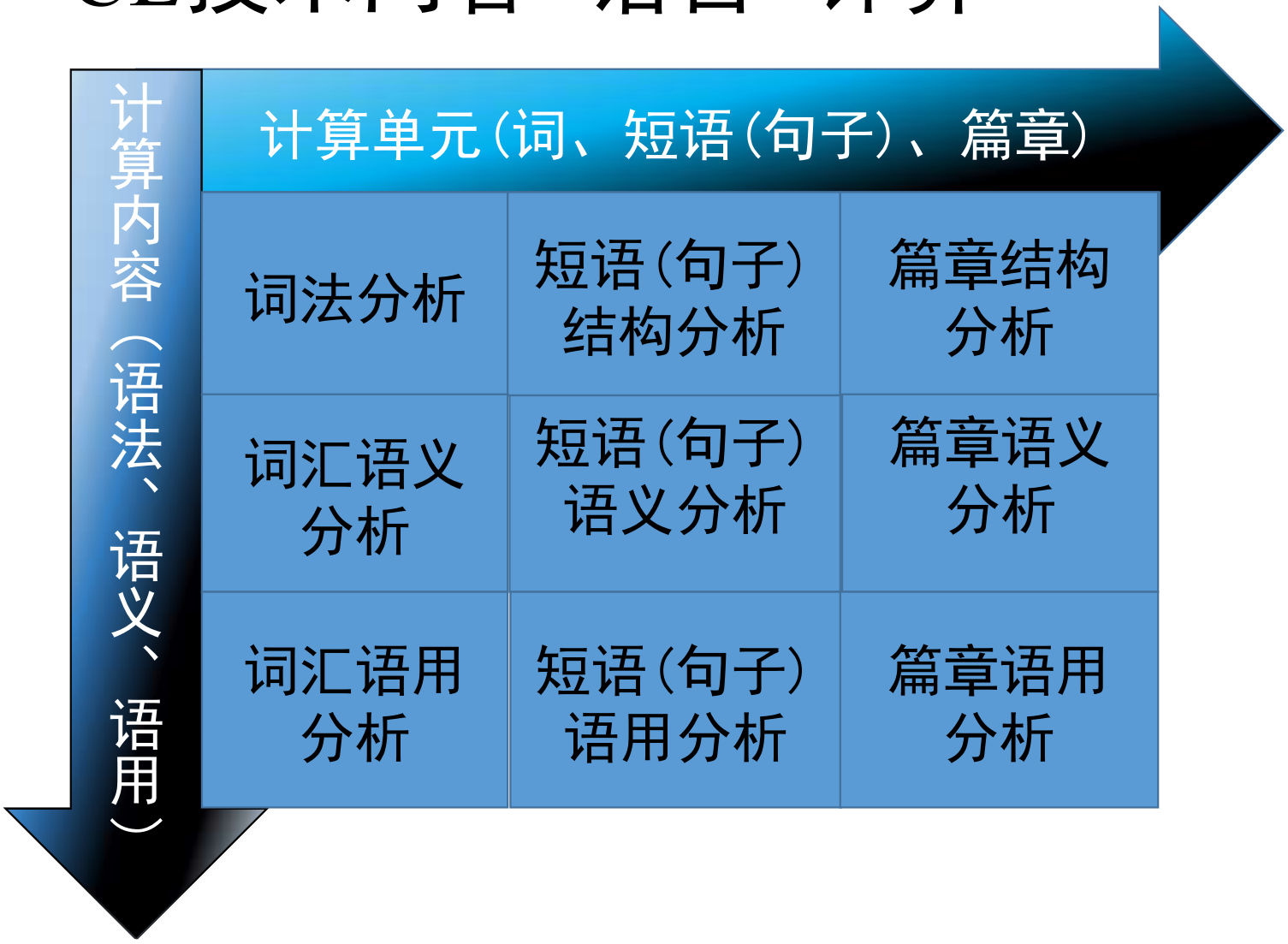
- Computational Linguistics(CL) 之别称
  - Natural Language Processing (NLP，自然语言处理)
  - Language Engineering (LE，语言工程)
  - Human Language Technology (HLT，人类语言技术)
  - Computer Speech and Language Processing
  - …
- Computation + Linguistics
- 在使用上各有侧重


计算机科学
计算语言学
语言学

CL技术内容=语言+



CL技术内容=语言+计算



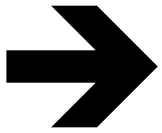
CL应用

- 拼写检查
- 语法检查
- 信息检索
- 文本摘要
- 机器翻译
- 问答系统
- 人机对话
- ……

CL研究的五个任务

- 提出语言问题：从语言中提出计算语言学要解决的问题
- 语言问题的数学建模：利用数学工具形式化语言问题
- 计算模型构建：为形式化的数学问题建立可计算的算法
- 编程实现：采用某种计算机语言实现算法并运行
- 评估：判断计算机算法对该语言学问题解决得如何

- 示例说明五个问题
  - 以汉语切分为例



提出语言问题

- 内塔尼亚胡说的话在美国会引起强烈反响.
  - 为理解该符号串的意思，需要首先知道该字串由哪些词组成
    - 内塔尼亚胡/说/的/话/在/美/国会/引起/强烈/反响.
    - 内塔尼亚胡/说/的/话/在/美国/会/引起/强烈/反响.
    - 内塔尼亚/胡说/的/话/在/美国/会/引起/强烈/反响.
    - 内塔尼亚/胡说/的/话/在/美/国会/引起/强烈/反响.
    - ……
- 语言问题：
  - 切分为词的时候出现多种不同的选择

语言问题的数学建模

- 一个简单的切分数学模型：M(F ,W ,T ,K)
  - F 是一个函数族, W是词典, T 是输入文本, K知识库.
  - 对任意 $t \in T$, 一个切分定义为一个函数：
  - $f(t \mid k)=w_1 w_2 \ldots w_n$,
  - 其中$f \in F$; $w_1 w_2 \ldots w_n \in W$; $k \in K$。

计算模型构建

- 计算模型
- 例如：切分作为函数来建模时，构造出这个函数
- 一些已有的算法：
  - 前向最大匹配算法(Forward Maximum Match algorithm)
    - 内塔尼亚胡/说/的/话/在/美国/会/引起/强烈/反响
  - 后向最大匹配算法(Backward Maximum Match algorithm)
    - 内/塔/尼/亚/胡说/的/话/在/美/国会/引起/强烈/反响
  - …

编程实现

- 算法的计算机语言实现
  - 选择编程语言
  - 指定一些参数
    - 例如，前向最大匹配要指定最大匹配长度参数
  - 设计一些数据结构
  - ……

评估

- 构造评估数据：测试语料(Corpus)
  - 切分任务：需要人工构建一些切分好的文本语料
  - 翻译任务：翻译参考答案1或多种
  - ……
- 设计评估准则
  - 比如对于切分任务：准确率、召回率、F1等
  - 翻译任务：BLEU值等
  - ……

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

为何研究计算语言学

- 技术应用方面
  - 使得人-机\人-人交互更自然、方便、快捷…
    - 基于自然语言的搜索引擎：信息获取便捷
    - 机器翻译：不同语种的人际交流更方便
    - 自动写作：用于新闻稿起草，提高工作效率
    - 人机对话：用于自动客服，降低服务成本
    - …

为何研究计算语言学

- 科学探索方面
  - 帮助揭示人类语言及语言处理的秘密
    - 人类语言处理是语言处理的仅有原型
  - 帮助揭示人类思维的本质
    - 语言是思维的外壳

为何研究计算语言学

- 社会价值
  - 基于自然语言的人机交互缓解甚至消除人对工具的异化
    - 异化：人类每发明一种新工具，在获得其所带来成果的同时，又总是使人成为工具的奴隶
    - 计算机：机器语言→编程语言→ 高级语言→……→NL

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

■ 理解如下语音输入

yi (3)	wei (4)	he (2)	dui (4)	ba (1)	ren (2)	zhen (4)	dong (4)	shou (3)

■音➔字

yi (3)	wei (4)	he (2)	dui (4)	ba (1)	ren (2)	zhen (4)	dong (4)	shou (3)
以								
乙								
已								
矣								
…								

■音➔字

yi (3)	wei (4)	he (2)	dui (4)	ba (1)	ren (2)	zhen (4)	dong (4)	shou (3)
以	为	何	对	巴	人	阵	动	手
乙	喂	和	队	八	任	朕	洞	首
已	位	河	兑	吧	仁	震	栋	守
矣	味	核		扒	壬	振	冻	狩
…	…	…		…	…	…	…	

■音➔字

yi (3)	wei (4)	he (2)	dui (4)	ba (1)	ren (2)	zhen (4)	dong (4)	shou (3)
以	为	何	对	巴	人	阵	动	手
乙	喂	和	队	八	任	朕	洞	首
已	位	河	兑	吧	仁	震	栋	守
矣	味	核		扒	壬	振	冻	狩
…	…	…		…	…	…	…	

- 字→词
  - 以为何对巴人阵动手
    - 以/为何/对/巴人/阵动/手
    - 以/为何/对/巴人/阵/动手
    - 以/为何/对/巴人阵/动手
    - 以为/何/对/巴人/阵动/手
    - 以为/何/对/巴人/阵/动手
    - 以为/何/对/巴人阵/动手
    - ……

■词性选择

以	为何	对	巴人阵	动手
动词	疑问词	动词	专有名词	动词
介词		副词		
专有名词		形容词		
		名词\量词		

- ■情况1：对=动词：以是对的主语，巴人阵是动手的主语
  - ■以为何是对的，巴人阵动手
- ■情况2：对=介词：以是动手的主语，巴人阵是介词宾语
  - ■以动手、动手对象是巴人阵

■词义选择

以	为何	对	巴人阵	动手
拿	表原因的疑问	回答	某机构	开始
用		针对		用手做
按照		面向		打人
以色列		适合		
		对偶词句		
		双		

- 句子意义选择
  - 以为何对巴人阵动手

- 句子的意义的深层理解：语用
  - 作者要通过这句话传达什么意思，需要语篇和更多的背景、专家知识帮助解读：语用歧义

CL: 核心问题

- 歧义(Ambiguity):
  - 包含了汉语的哪些音：语音
  - 用了汉语中的那些词：词汇
  - 词用了哪种句法属性：词性
  - 词采用了哪个义项：词义
  - 词是如何构成一个句子：句法
  - 句子的意义是什么：句义
  - 这句话用来干什么：语用

- NLP核心问题：歧义(Ambiguity)
- NLP核心任务：消歧(Ambiguity resolution)
- 如何消歧？

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

计算语言学：基于知识的活动

- 语言处理需要的知识
  - 语言学知识
    - 语音知识：yi(3)音有几个字“以、已”对应、…
    - 词汇知识：“以后”是词、“已后”不是词、……
    - 句法知识：“看书”是短语，“他书” 不是、…
    - 语义知识：看有几个不同的义项：视线接触、阅读、…
    - 语用知识：…

计算语言学：基于知识的活动

- 语言处理需要的知识
  - 非语言学的知识
    - 常识知识(感知经验)
      - 水是透明的、石头是硬的、…
      - 咬了猎人的狗　　vs　　骂了猎人的狗
    - 世界知识：中国位于亚洲、…
    - 文化知识：各民族、地域的不同文化风俗…

获取和表示知识

- 来源于人类理性 (理性主义)
  - 人们总结经验、专家创建知识
  - 以规则、谓词等各种知识表示方法进行表示
  - 例如：
  - 语言学知识
    - S→NP VP
  - 非语言学知识
    - $\forall x\ p(x), x \in$ 所有生物构成的集合，p(x):x会死的

获取和表示知识

- 来源于语料(corpus) (经验主义)
  - 未标语料、标注语料
  - 计算机从语料中自动学习、抽取知识
  - 相比于理性主义知识，更为细粒度、更多量化
  - 语言学知识
    - $p(is|it) = 0.003$
  - 非语言学知识
    - $p(Watson, CEO, IBM) = 0.05$

表示和使用知识：知识形式化、算法化

- 模型：语言知识的形式化表示和运算
  - 状态机 (FSA, FST,…)
  - 形式语言系统 (CFG, UG,…)
  - 逻辑系统 (Predicate calculus, …)
  - 概率统计模型 (LR、MC, …)
  - 神经网络模型(SGNS, RNN, Transformer, BERT、GPT…)
- 算法：形式化的语言知识的计算操作
  - 搜索、动态规划
  - 各种机器学习\深度学习算法

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

1940s-1950s：奠基

- 自动机
  - Turing 1936：计算、Turing机
  - McCulloch-Pitts (1943)：神经元
  - Kleene (1951/1956)：正则表达式
  - Shannon (1948)：自动机与马尔科夫模型的关联
  - Chomsky (1956)/Backus(1959)/Naur(1960): CFG
- 概率/信息理论
  - Shannon (1948)：信息熵
  - Bell Labs (1952)：语音识别

1957-1970：两个阵营

- 符号学派
  - Zellig Harris 1958: 级联有限状态转录机(FST)
  - Chomsky：短语结构语法(PSG)
  - Newell&Simon: 逻辑理论家，通用问题求解
- 统计学派
  - Bledsoe and Browning (1959): Bayesian OCR
  - Denes (1959): 结合语法和声学概率的ASR

1970-1983：4个范式:

- 随机、统计方法
  - 隐马尔科夫模型(HMM 1972)
- 逻辑、形式方法
  - 定子句语法 (DCG，Pereira & Warren 1980)
  - 功能语法(FG, Kay 1979), 合一语法(UG, Bresnan & Kaplan 1982)
- 人工智能方法(认知分析)
  - 积木世界(Shrdlu, Winograd 1972)
  - 框架、脚本(Schank and Abelson1977)：
- 语言建模方法(语言分析)
  - 语篇结构、焦点(Grosz et al)

1983-1993：有限状态和概率模型的回归:

- 有限状态模型
  - Kaplan and Kay (1981): Phonology/Morphology
  - Church (1980): Syntax
- 概率模型
  - 为语言任务创建语料
  - NLP应用的早期统计版本
  - 方法上进一步明晰
    - Can’t test your hypothesis on the data you used to build it!
    - Training sets and test sets

1994-2003：SML的方法崛起

- 简单统计模型广泛使用
  - ACL conference:
    - 1990: 39 articles 1 statistical
    - 2003 62 articles 48 statistical
- 更为复杂的统计模型(统计机器学习模型：SML)
  - ME\MEMM\SVM\CRF\LDA……

2004-2013：方法融合、面向应用

- 规则与统计方法的不断融合
  - 自动机+统计、逻辑+概率、…
- 面向应用
  - Web based :
    - IR/IE meets NLP
    - ….
  - Robot based :
    - ASR\TTS\Dialogue meets NLP

2014-：深层神经网络

- 词表示模型：SGNS\GloVe/Skip-Thought/…
- 序列编码模型：RNN(LSTM/GRU)
- 序列转换模型：各种序列编解码组合(Seq2Seq)
- 注意力模型：soft/hard/self/hierarchical…
- 记忆网络/对抗生成网络/…
- Elmo/Transformer/BERT/GPT-3…
- 语言处理的预训练范式

- 代表性应用成果
  - 搜索：http://www.wolframalpha.com/
  - 问答系统：IBM Watson
  - 翻译：http://translate.google.cn/
  - 推荐：Amazon图书推荐
  - 基于GPT-3的写作

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

教材

- 【JMBook】Daniel Jurafsky and James H. Martin,2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall. 2nd Edition
- 3rd进行中，可参考

■ 课程进度安排

周次	日期	阅读章节 [JMBook]	上课内容	作业
03	09.18	Chap 1	课程介绍	
04	09.27	Chap 2 , 3	英语词法分析：RegExp/FSA/FST	教材阅读报告1
05	10.02		国庆停课1次	
06	10.09		汉语词法分析1/语言模型1	
07	10.16	Chap 4, 16	语言模型2/PP/神经语言模型	教材阅读报告2
08	10.23		汉语词法分析2/NER	
09	11.30	Chap 5, 6	POS tagging：序列标注	教材阅读报告3
10	11.06	Chap 12,13	句法分析：CFG/CKY/Chunking	教材阅读报告4
11	11.13	Chap 14	统计句法分析：PCFG/PCKY/DP	教材阅读报告5
12	11.20		词义分析：词表示	编程作业？
13	11.27		句义分析：句子表示/SRL	
14	12.04		语篇分析：概述/LDA	
15	12.11		应用：人机对话	
16	12.18		课堂测试	

- 课程基础要求
  - 形式语言与自动机：有限状态自动机、正则语言…
  - 概率统计：条件概率、极大似然估计…
  - 信息论：熵、交叉熵…
  - 最优化：动态规划、梯度下降…
  - Python语言编程

读物

- Conference
  - ACL\COLING\EMNLP\…
  - SIGHAN\SIGDIAL\...
  - CCL\NLPCC\...
  - IJCAI\AAAI\...
- http://www.aclweb.org/anthology/

大纲

- 计算语言学是什么？
- 为何研究计算语言学？
- 计算语言学的关键问题是什么？
- 解决计算语言学关键问题依靠什么？
- 计算语言学已经有什么？
- 教材和读物
- 课程评估方法

- 教材阅读报告(5次)：30%
  - 提交截至时间：对应内容上课周的周五24:00时
  - 提交内容：所阅读章节的内容综述、存在的问题等
- 编程作业(1次)：30%
  - 提交截至时间：12月20日24:00时
  - 提交内容：说明文档(算法、评测等描述)、python代码等
- 随堂测试(1次)：40%
  - 最后一次课随堂测试：当时提交

交作业邮箱: colingrad@163.com

邮件请注明姓名学号,不接受延迟提交！

Thank you !