

第一讲 绪论

<http://ccl.pku.edu.cn/alcourse/nlp>



自然语言处理 (NLP)
计算语言学 (CL)

- 1. 研究内容
- 2. 研究方式
- 3. 应用领域
- 4. 发展简史



0 引子

我们可以期待，总有一天机器会同人在所有的智能领域里竞争起来。但是，如何开始呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，还有一种办法也应加以考虑，就是为机器配备具有智能的、可用钱买到的意识器官，然后，教这种机器理解并且说英语。这个过程可以仿效通常小孩子学话的方式进行。我不能确定到底哪个出发点更好，但应该都值得一试。

—— A. M. Turing, *Computing Machinery and Intelligence*,
Mind Vol.59, 1950.



交谈

甲：听说吴妈打赢了阿Q。

乙：不错，阿Q确实被吴妈打败了。

甲：这个结果有些出人意料。

乙：阿Q是大意失荆州，怪不得别人。



图灵测试 (Turing Test)

- 一个封闭的小屋里面有两间房A和B，其中一间房中是人，一间是机器
- 屋外有一个人H
- H向A，B房内的实体提出各种问题
- H根据来自A，B的回答判断A，B中谁是计算机，谁是人
- 如果H无法判断，则证明计算机具有人的智能



Eliza 问答系统

(by Joseph Weizenbaum, MIT, 1960s)

Young woman: Men are all alike.

Eliza: In what way?

Young woman: They're always bugging us about something specific or other.

Eliza: Can you think of a specific example?

Young woman: Well, my boyfriend made me come here.

Eliza: Your boyfriend made you come here?

Young woman: He says I'm depressed much of the time.

Eliza: I am sorry to hear you are depressed.

Young woman: It's true. I'm unhappy.

Eliza: Do you think coming here will help you not to be unhappy?

Young woman: I need some help, that much seems certain.

Eliza: What would it mean to you if you got some help?

Young woman: Perhaps I could learn to get along with my mother.

Eliza: Tell me more about your family.

Young woman: My mother takes care of me.

Eliza: Who else in your family takes care of you?

Young woman: My father.

Eliza: Your father.

Young woman: You are like my father in some ways.

Eliza: What resemblance do you see?

Joseph Weizenbaum, 1966, **ELIZA — a computer program for the study of natural language communication between man and machine**, *Computational Linguistics*, Jan. 1966
Vol.9, No.1



Searle的“Chinese Room”

1980s

哲学家约翰·西尔勒(John R. Searle)否认可以模拟大脑算法的计算机具有意识，他在上个世纪80年代给出了一个著名的理想实验“西尔勒的中文屋子”来支持他的论点。

实验（改编版）的要旨是这样的：

一个用中文讲的笑话，中国人听了会哈哈大笑。如果用一台可以执行复杂算法的计算机根据这个笑话的词句进行一系列运算，得出的结果正确，也是“哈哈大笑”，于是计算机用它自己的方式笑了几声，我们是否可以认为这台计算机“理解”了这个笑话呢？

类似的，用一大群不懂中文的美国人取代计算机，他们拿着铅笔和纸重复计算机所做的一切，因为算法很复杂，可能要全美不懂中文的美国人算上一年才得到了结果“哈哈大笑”，他们派一个代表出来笑了几声。虽然反应很慢，但他们和一个中国人做得一样好，不过，这样仍然无法认定这群美国人“理解”了这个中文笑话。

Searle, John. R. (1980) Minds, brains, and programs. In *Behavioral and Brain Sciences* 3 (3): 417-457



Watson问答系统 (Feb. 2011)

- Q: Each year the EU selects capitals of culture: one of the 2010 cities was this Turkish "meeting place of cultures"
- A: Istanbul (Watson's first answer with 81% confidence)

- Q: Elected every 5 years, it has 736 members from 7 parties
- A: parliament (Watson's first answer with 66% confidence)

- Q: Dialects of this language include Wu, Yue & Hakka
- A: Chinese (Watson's answer: Cantonese 41% confidence)

.....

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, Chris Welty, 2010, **Building Watson: An Overview of the DeepQA Project**. *AI Magazine*, 2010. Vol.31, No.3.



关于语言，我们了解多少？

- (1) 人用来交际的“语言”具有什么样的性质？这些性质又是如何影响交际过程的？
- (2) 人用来交际的“语言”跟机器可以“理解”的语言有什么样的关系（相同和不同）？
- (3) 人是如何运用“语言”进行交际的？
- (4) 人运用语言进行交际的过程是否可以描述为一个机械的过程？
- (5) 什么叫做“理解”一种语言？

.....



机器语言 vs. 自然语言

```
#include <stdio.h>

void main ()
{
    int x=1, y=2, z;
    z = x+++y;
    z = x+++++y;
    z = x++ + ++y;
    z = x+++ +y;
    z = x + (++y);

    printf("z=%d\n",z);
}
```

人们以为他对她有“意思”，于是，建议他对她“意思意思”。他说，他没那种“意思”。她则反问，你们是什么“意思”。大伙中有的觉得很有“意思”，有的则认为真没“意思”。



定义：什么是NLP？

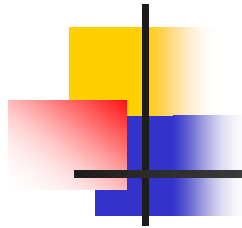
- 用计算机来分析和生成自然语言（文本、语音）
- 目的是让人类可以用自然语言形式跟计算机系统进行人机交互，从而更便捷、有效地进行信息管理
- NLP的具体表现形式包括机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等等。

H. Cunningham. 1999, **A Definition and Short History of Language Engineering**. *Journal of Natural Language Engineering*, pp. 1--16, vol 5, 1999.



1 NLP的研究内容

- 从计算的角度来研究语言的性质
- 将语言作为计算对象来研究相应的算法



1.1 从计算角度研究语言

所谓从计算的角度来看语言的性质，就是要求将人们对语言的结构规律的认识以精确的、形式化的、可计算的方式呈现出来，而不是像其他语言学研究那样，在表述语言的结构规律时一般采用非形式化的表达形式。



例子

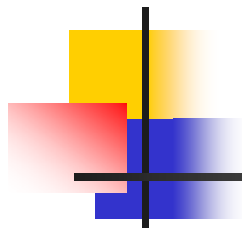
- 1a 张三已经赶跑了李四
- 1b 张三已经把李四赶跑了
- 1c 李四已经被张三赶跑了

- 2a 吴妈以前很喜欢阿Q的理论
- 2b * 吴妈以前把阿Q的理论很喜欢
- 2c * 阿Q的理论以前被吴妈很喜欢



语法规律

- 1) 汉语中的一个基本句型是:
 $P_0: X + \text{动词} + Y$
- 2) P_0 可以变换为“把”字句或“被”字句
 $P_1: X + \text{把} + Y + \text{动词}$
 $P_2: Y + \text{被} + X + \text{动词}$
- 3) P_0 满足条件 C_1, C_2, \dots, C_i 时可以变换为 P_1, P_2 ;
否则, 不可以变换为 P_1, P_2 ;



1.2 将语言作为计算对象

所谓将语言作为计算对象来研究相应的**算法**，是研究如何以机械的、规定了严格操作步骤的程序来处理语言对象（主要是自然语言对象，当然也可以是形式语言对象），包括一个语言片断（比如词组、句子或篇章）中大小语言单位的识别，该语言片断的结构和意义的分析（自然语言理解），以及如何生成一个语言片断来表达确定的意思（自然语言生成），等等

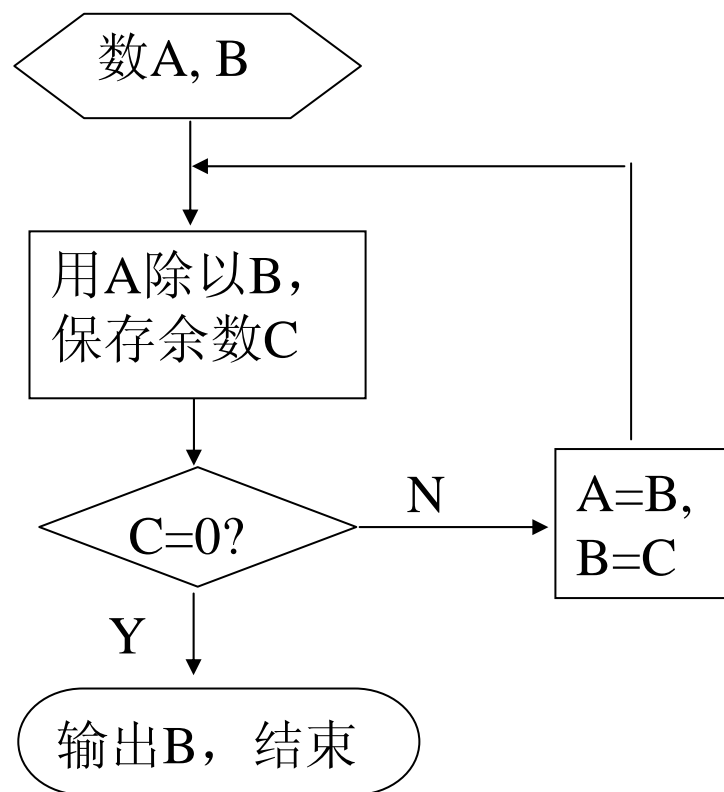


算法 (algorithm)

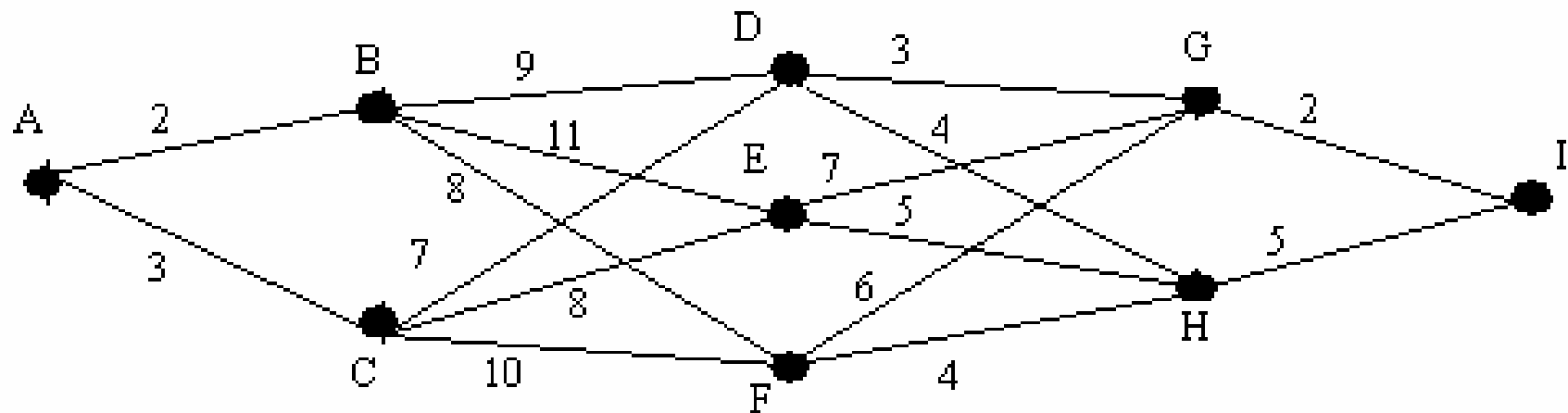
- (1) 通用性：算法是针对一类问题的，而不仅仅是用于解决某一个具体问题。
- (2) 机械性：算法的每一个步骤都是机械的，确定的。
- (3) 有限性：算法必须在有限步内结束。
- (4) 离散性：算法的输入数据及输出数据都是离散的符号。



算法实例（1）：求最大公约数



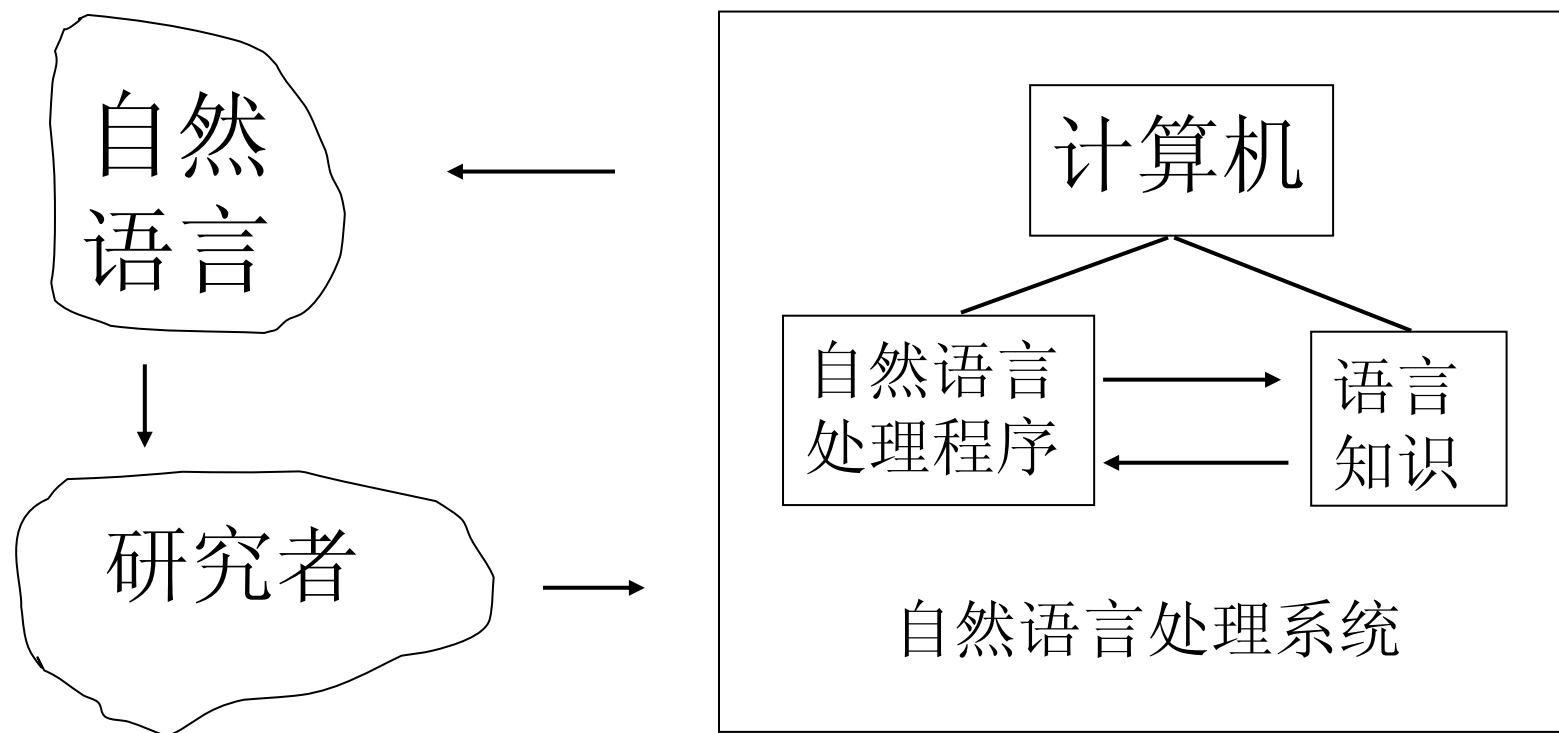
算法实例（2）：求最佳路径

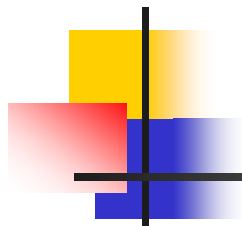


AB=2 AC=3 BD=9 BE=11 BF=8 CD=7 CE=8 CF=10
DG=3 DH=4 EG=7 EH=5 FG=6 FH=4 GI=2 HI=5

从A 到 I 的最佳路径是哪一条？

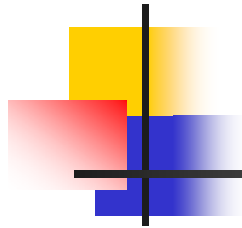
2 NLP的研究方式





动态视角（流程）

- S1: 研究者以特定的方式对自然语言 (NL_0) 的规律进行抽象, 以计算机能够处理的形式来表述关于自然语言的规律——得到语言知识K;
- S2: 针对特定的语言知识表示形式, 研制适合的分析 and 处理算法;
- S3: 根据算法编制计算机可执行的自然语言处理程序P。这样的程序加上语言知识, 加上计算机硬件系统, 共同构成一个自然语言处理系统 (NLPs) ;
- S4: 用这样一个自然语言处理系统对自然语言 NL_0 进行分析处理, 根据反馈的结果调整原来的设计, 改进NLPs。

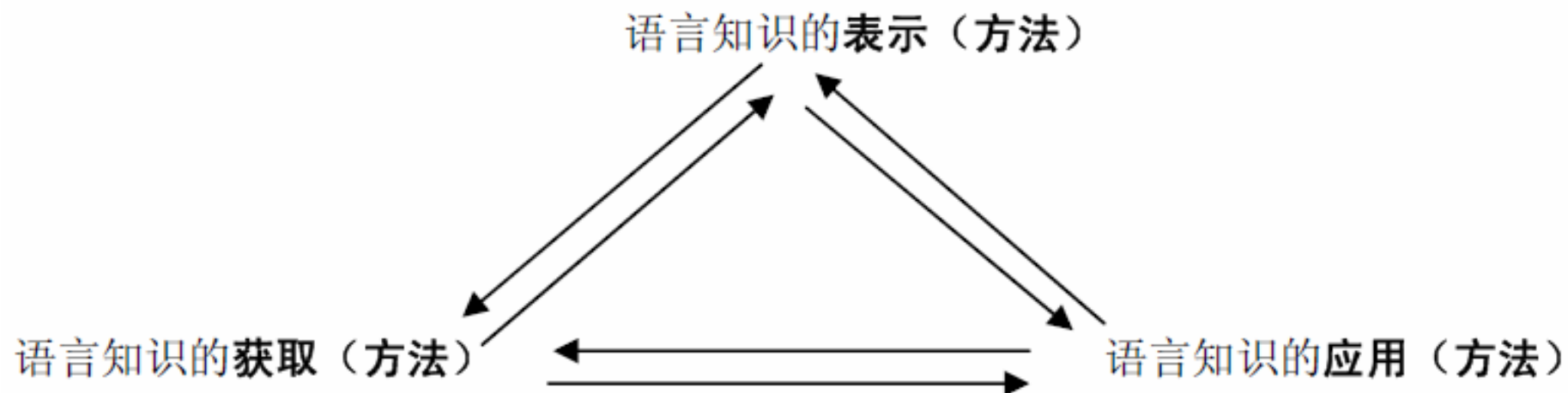


静态视角（模块）

- 语言对象 语音 字 词 词组 句子 篇章
- 语言知识 音系 形态 句法 语义 语篇
- 处理程序 stemmer annotator parser translator ...

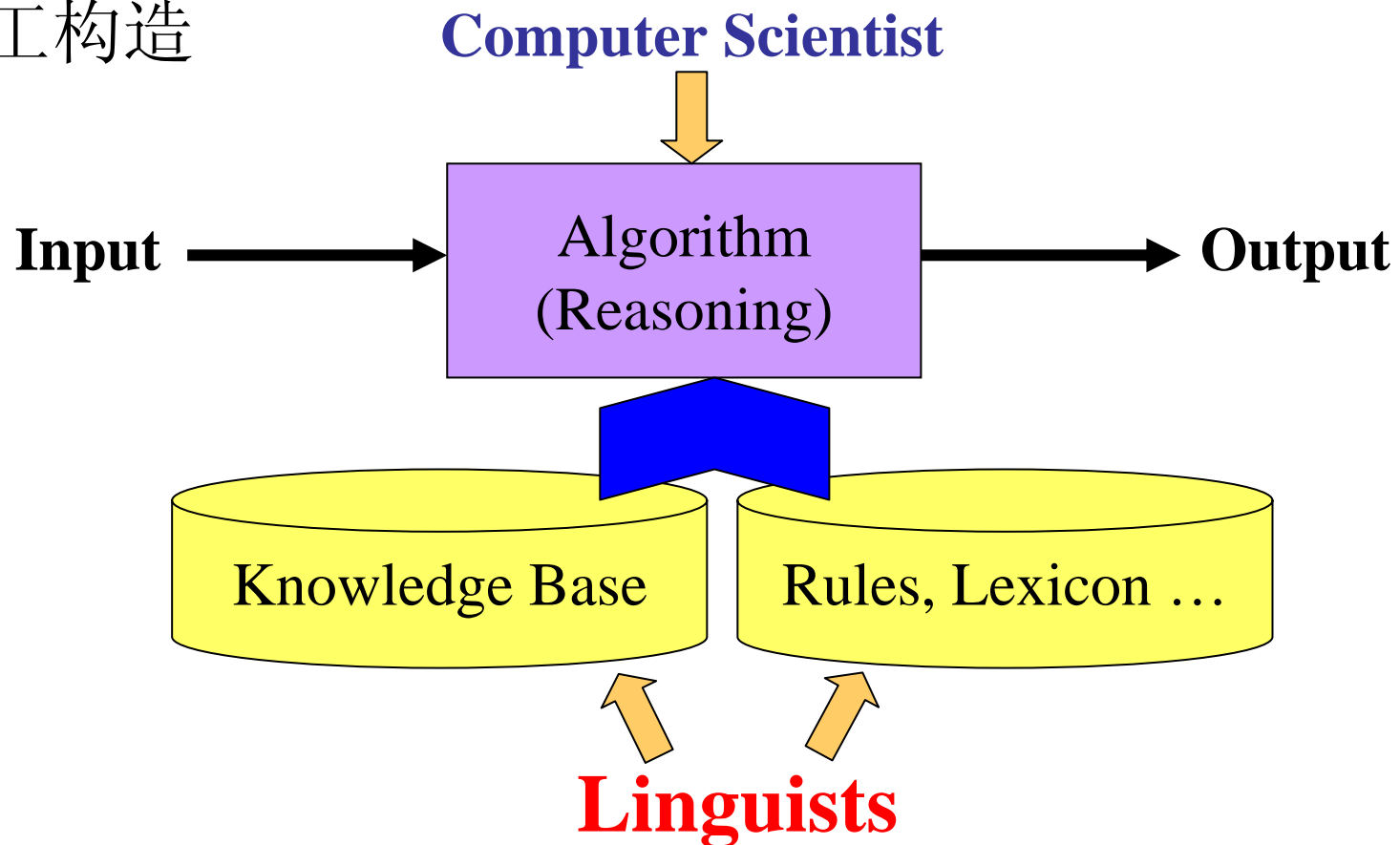


围绕“语言知识”看NLP

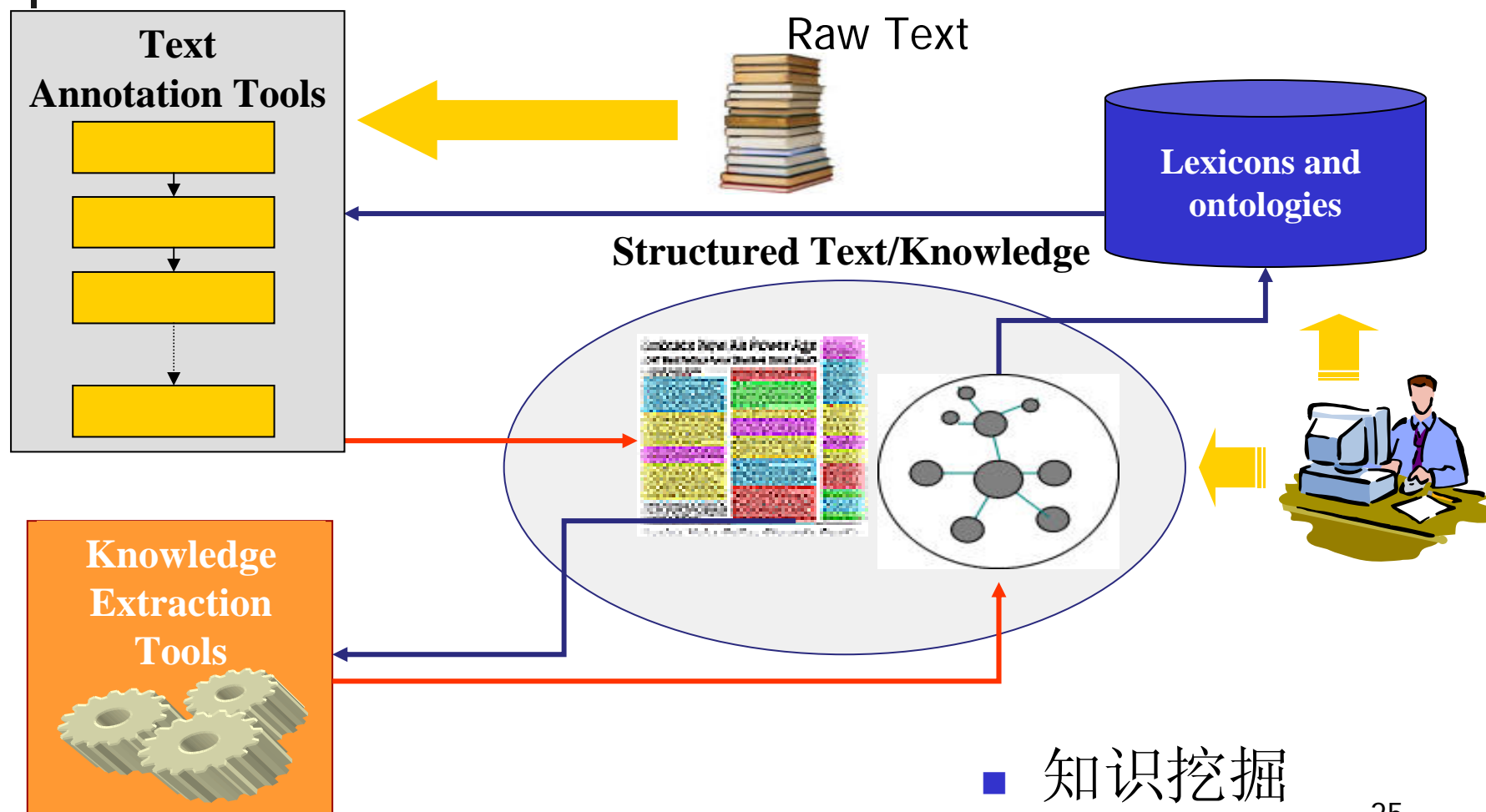


语言知识的获取方式 (1)

- 人工构造

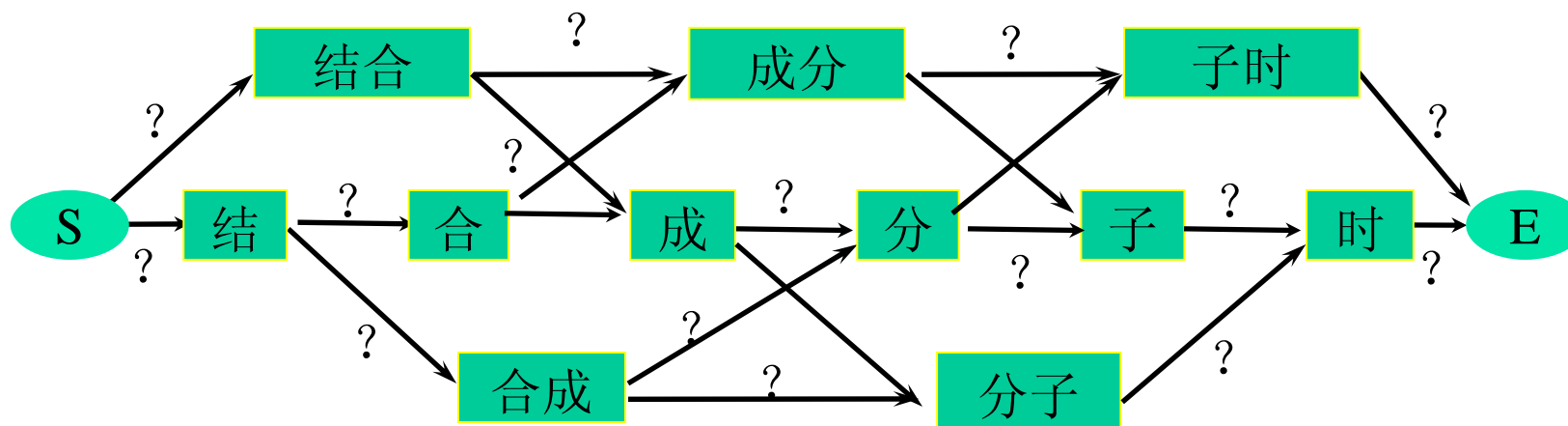


语言知识的获取方式 (2)

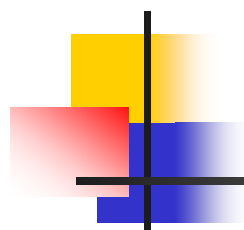


■ 知识挖掘

举例：“结合成分分子时”需要的分词知识



1	结/ 合/ 成/ 分/ 子/ 时	8	结/ 合成/ 分子/ 时
2	结/ 合/ 成/ 分/ 子时	9	结合/ 成/ 分/ 子/ 时
3	结/ 合/ 成/ 分子/ 时	10	结合/ 成/ 分/ 子时
4	结/ 合/ 成分/ 子/ 时	11	结合/ 成/ 分子/ 时
5	结/ 合/ 成分/ 子时	12	结合/ 成分/ 子/ 时
6	结/ 合成/ 分/ 子/ 时	13	结合/ 成分/ 子时
7	结/ 合成/ 分/ 子时		

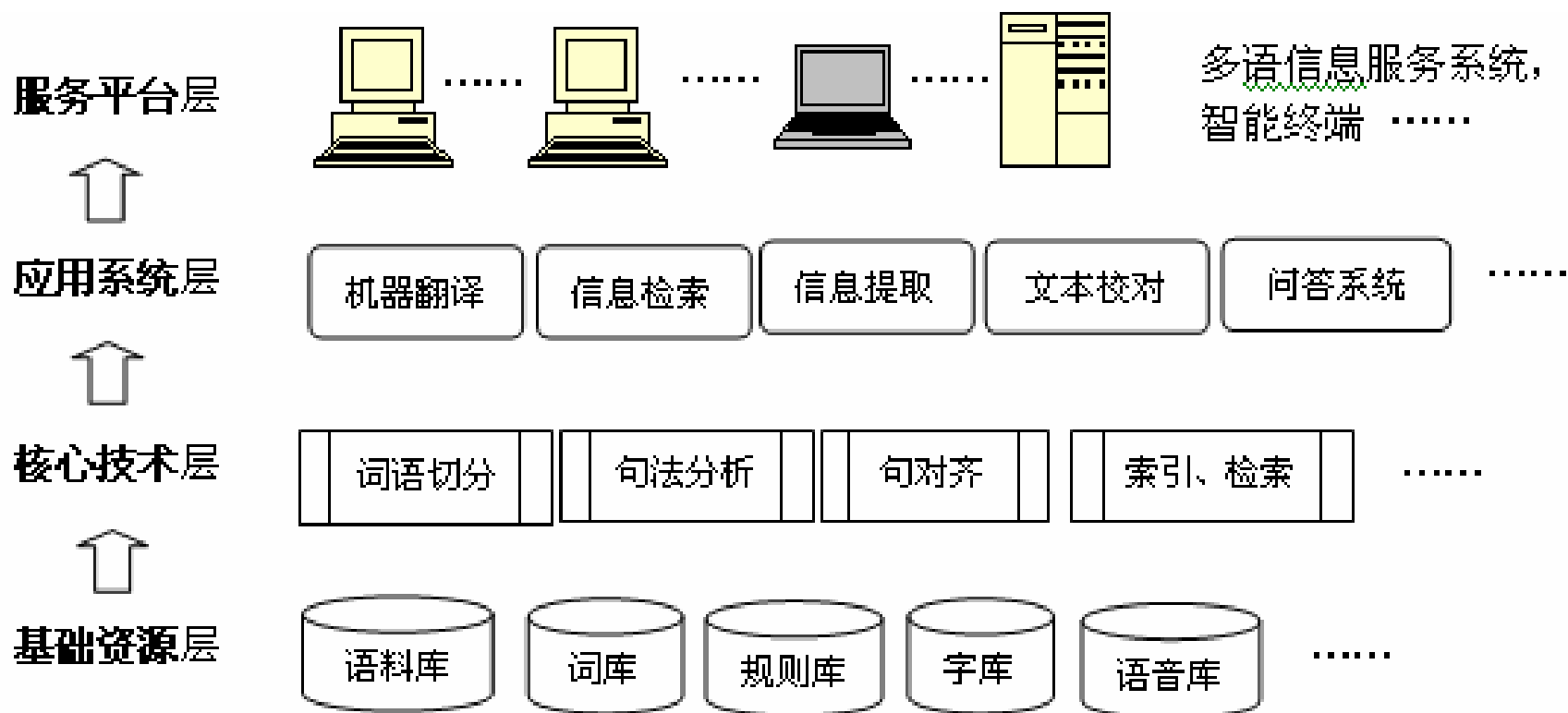


自然语言处理的宏观架构（1）

自然语言信息处理的对象、层次和任务

对象 任务 层次	书面文本 [视觉符号]		口语语音 [听觉符号]	
	文本理解 文本生成	[机器翻译 信息检索...] [文本摘要 问答系统...]	语音识别 语音合成	[口语翻译...] [口语问答...]
处理符号的意义				
处理符号的形式	汉字输入、存储、输出 篇章版式分解与生成		语音信号采集、 波形特征抽取、波形生成	

自然语言处理的宏观架构（2）





3 NLP的应用领域

- 机器翻译 (Machine Translation)
- 文本分类 (Text Classification)
- 信息检索 (Information Retrieval)
- 信息提取 (Information Extraction)
- 语音合成 (Speech Synthesis)
- 语音识别 (Speech Recognition)
- 人机接口 (Human-Machine Interface)
-



整句输入

- 转换

- 拼音输入：自动将拼音序列转化为汉字序列；

一个例子：

拼音输入：Ji qi fan yi ji qi ying yong ji qi kun nan

汉字序列：.....?

- 语音输入：自动将连续读音转化为汉字序列；
- 文语转换：将连续文本转化为语音信号输出；

- 校对

- 拼写校对：我们要京城(精诚)合作
- 文法检查。



文本检索 (Text Retrieval)

- 在 Internet 或数字图书馆上

- 输入词、短语或句子
- 检索相应的文档
- 例子:

和服

Search

- Question

- 如何得到想得到的结果?



未经中文分词处理时的检索结果

1.

电信运营商**和服务**提供商

采用奥维通的移动WiMAX解决方案,运营商和服务提供 商可以提供各种个人宽带服务

2.

关于做好党员联系**和服务**群众工作的意见

做好党员联系和服务群众工作,要以马克思列宁主义、毛泽东思想、邓小平理论和“三个代表”重要.....

3.

Guangzhou bomei leather co.,ltd

站长信息**和服务**中心:斗破苍穹 阴阳冕 九鼎记 凡人修仙传 猎国 九转金身决.....

4.

关于商品**和服务**实行明码标价的规定

根据《中华人民共和国价格法》修订的《关于商品**和服务**实行明码标价的规定》,

5.

Technical Support

利盟中国面向行业,办公和家庭提供彩色激光,黑白激光,喷墨,和多功能一体打印机及相关耗材**和服务**,是业届领先的打印解决方案的开发制造商。

.....

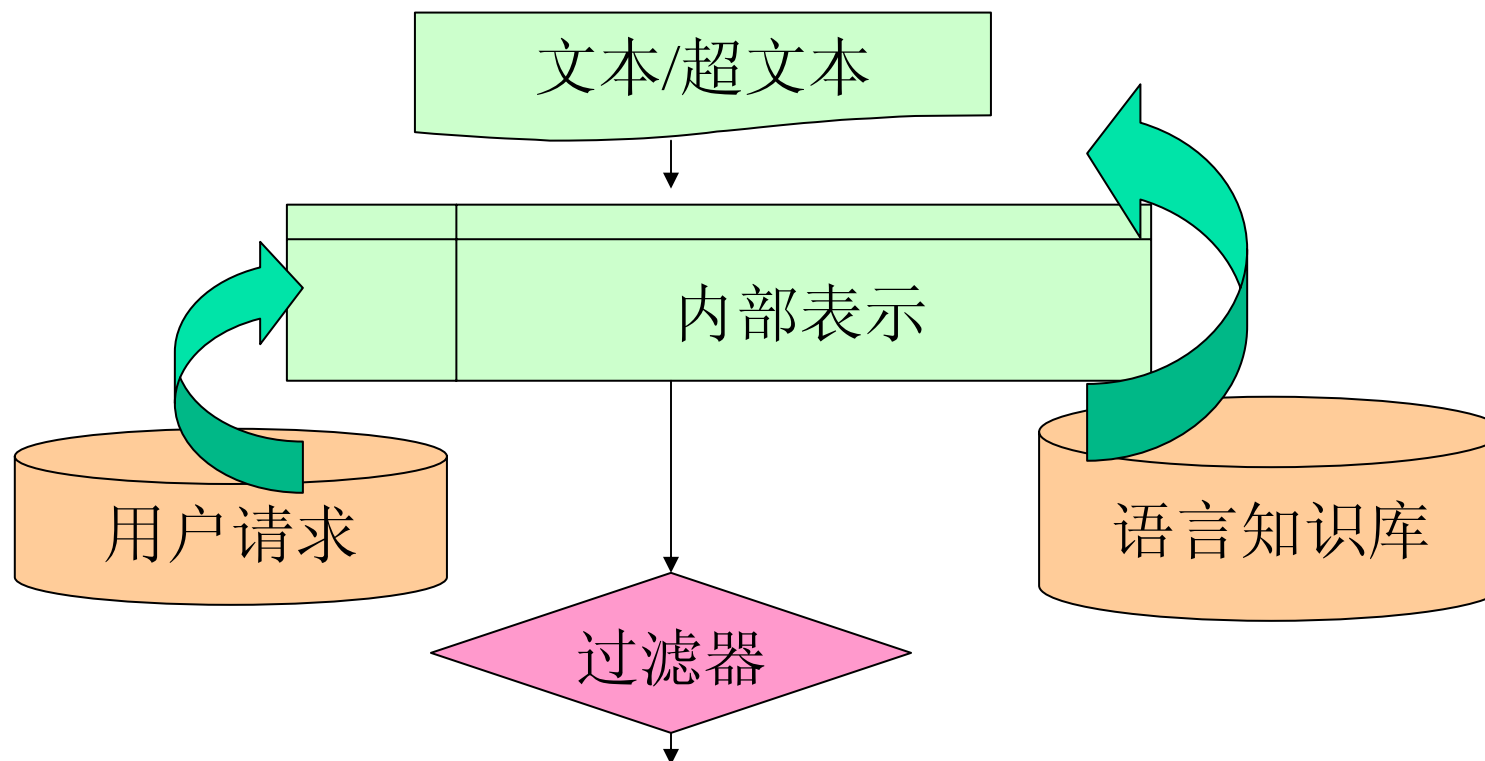


自动问答

- 简单问答
 - 问路线：颐和园怎么走？
 - 问天气：香港明天的天气如何？
 - 问航班/火车时刻
 - 网上购物问价格
 - 网上找人
- 问题描述（语言的表述）
- 答案返回（需要带摘要吗？）

文本信息过滤

- 例子：过滤色情网站（或网页）



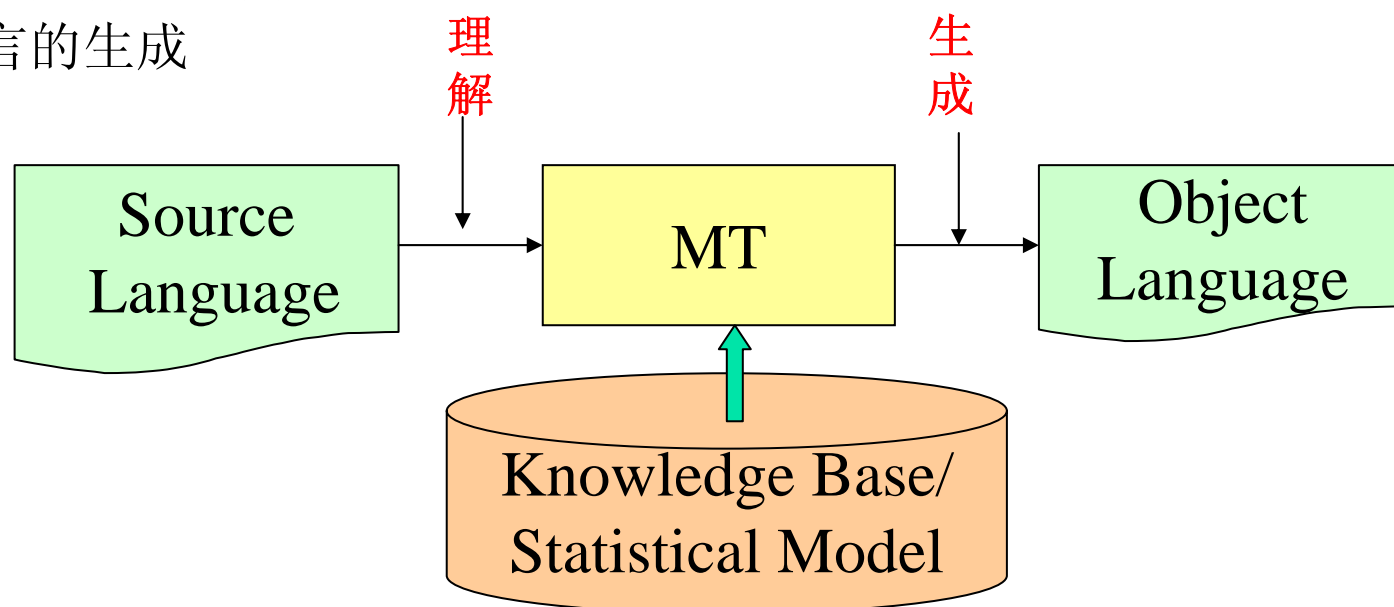


情感&观点分析

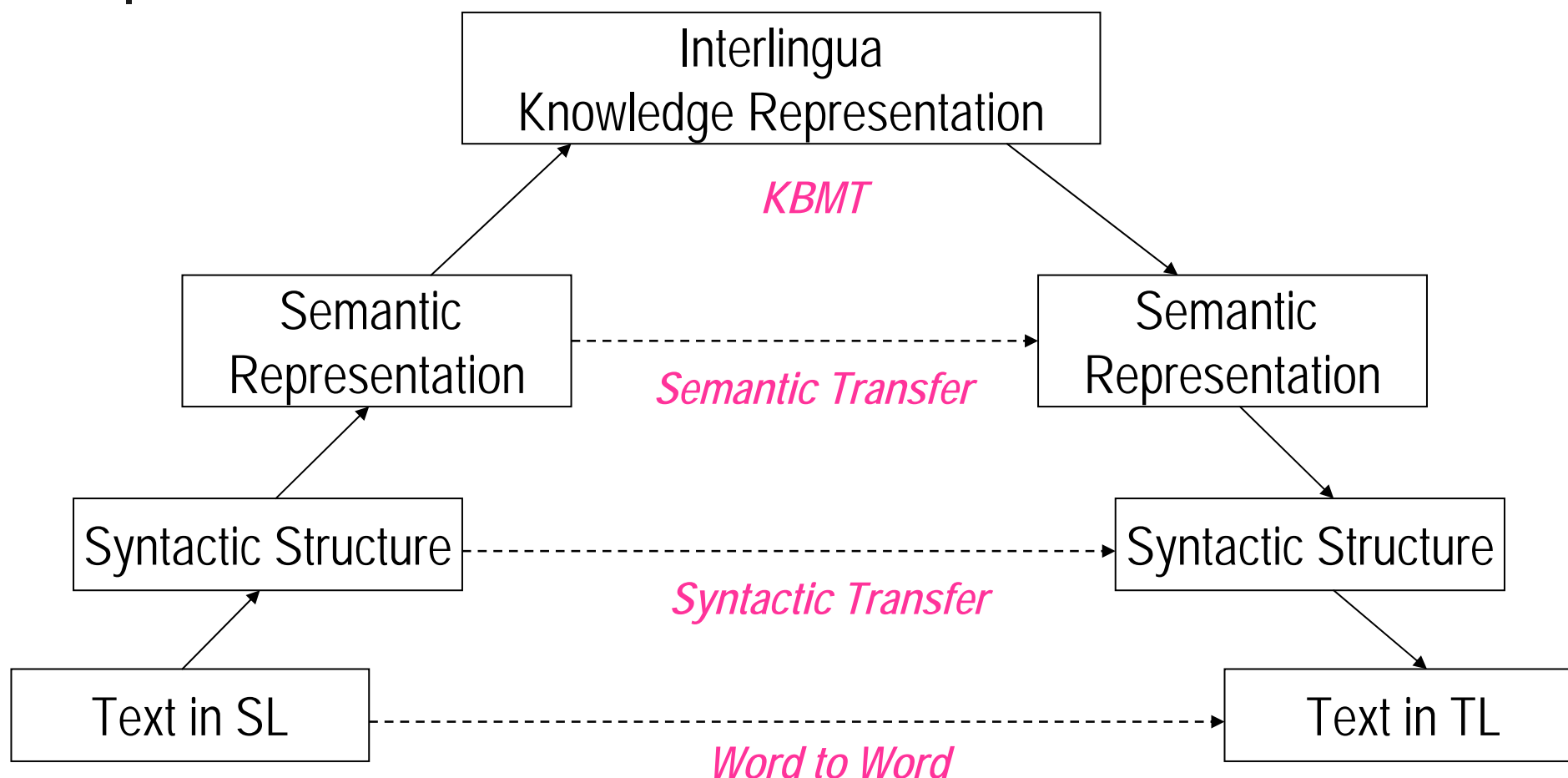
- 为什么要对文本进行情感分析？
 - 文本是人写的，必然带有人的感情和观点
 - 大量的应用需要情感与观点分析：
 - 大量的评论性文本：商品评论，服务质量评论，电影评论；
 - 带政治色彩的评论：敌对势力的攻击，法轮功的攻击，...
 - 无所不在的社会舆论：带有大量的观点或感情色彩；
- 情感与观点分析需要做什么？
 - 观点是什么？带有怎样的情感色彩（正面 / 负面）？
 - 是谁（Holder）发表的观点或表达的情感？
 - 针对的问题（对象）是什么？
 - 上述问题都需要通过文本分析提炼出来。

机器翻译

- 自动翻译：英语 \leftrightarrow 汉语
- 跨语言检索：输入汉语检索条目，返回满足要求的其他语言信息（检索条目的翻译）
- 两个过程：
 - 原语言的分析 and 理解
 - 目标语言的生成



机器翻译





机器翻译的困难

原文: The spirit is willing, but the flesh is weak

机器译文: 酒是好的, 但肉是馊的

正确译文: 心有余而力不足

原文: How are you? 机器译文: 怎么是你?

原文: How old are you? 机器译文: 怎么老是你?



4 NLP的发展简史

- 1950 — 1960年代 Warren Weaver(1949)
Turing Test(1950)
- 1960 — 1970年代 The first MTs(1954)
ALPAC(1964-1966)
- 1970 — 1990年代 Searle's Chinese Room(1980)
The first PC version of MTs(early 1980s)
- 1990 — 至今 MT is available on the Web(1994)
.....

1. [http://ccl.pku.edu.cn/doubtfire/NLP/Machine Translation/overview/Machine Translation Past and Future.htm](http://ccl.pku.edu.cn/doubtfire/NLP/Machine%20Translation/overview/Machine%20Translation%20Past%20and%20Future.htm)
2. W.John Hutchins, 2001, Machine translation over fifty years, In *Histoire, Epistemologie, Langage*, Tome XXII, fasc. 1 (2001), p.7-31
3. 冯志伟 (2004) 《机器翻译研究》第一章1.3节, 中国对外翻译出版公司。



萌芽期(1950s及之前)

- 1933: 法国的 Georges Artsrouni & 俄国的 Peter Trojanskij建议:构建机器多语言词典;
- 1946-1947:美国的Andrew Booth 和 Warren Weaver,提出了机器翻译的设想.
- 1950s: Yehoshua Bar-Hillel(MIT): 1952年举办了 1st MT会议, 会上, Leon Dostert(Georgetown Univ.)建议开发演示系统, 以吸引基金的投资.
- 1955年, 第一个演示系统在 IBM & Georgetown 开发, 包含250 个词和 6 条句法规则,实现 Russia — English;
- 第一本期刊: Mechanical Translation(1953-1970) 在 MIT出版.
- 第1篇博士论文 1953在MIT由 Anthony G. Oettinger完成: 俄语机器词典.



理论基础（1950s）

- 形式语言 (Chomsky, Kleene, Backus).
 - 语言描述的形式化：对语言按复杂程度分类，对不同类语言进行形式化描述；
 - 语言处理的形式化：对不同类语言进行自动识别和分析
- 概率与信息论方法
 - 语言的理解作为解码：Shannon的噪音管道模型；
 - 使用概率方法：将语言的产生看成随机过程；



低谷期：ALPAC 报告(1966)

- Automatic Language Processing Advisory Committee (ALPAC) (1964, USA)
- ALPAC 报告的内容 (1966) :
 - “There is no immediate or predictable prospects of useful machine translation”—— Ends funding MT.
 - Only support fundamental research in CL
- ALPAC 报告的原因:
 - 遇到了语义障碍 (Semantic Barrier)
 - Bar-Hillel的批评: 需要真实世界的知识



恢复期（1970s）

- 1971: W. Woods: Lunar, IR(ATN)
- 1972: T. Winograd: SHRDLU (Lisp)
- 1973: Schank: Concept Dependency (CD Theory)
 - MARGIE 1975: (Meaning, Analysis, Response Generation and Inference on English) on CD: NLU



发展期（1980s）

- 强调知识的重要性
 - 人工智能的发展
 - 日本的第5代计算机（知识处理系统）
 - 语言知识与语言分析的分离
- 语义知识的表示与知识库构建：CYC，Wordnet等；
- 机器翻译在受限领域获得成功：加拿大Meteo；
- 句法语义和词汇语义理论的蓬勃发展：
 - GPSG (Generalized PSG: Gazdar),
 - GB (Governing and Binding: N. Chomsky) ,
 - FUG (Functional Unification Grammar: M. Kay)
 - ...



经验方法再度受重视（1990s）

- IBM's P.Brown(1988-2nd TMI—Theoretical & Methodological Issues in MT, 1990-CL—Computational Linguistics, 1993-CL): 倡导机器翻译的统计方法;
- 基于统计与语料库的方法逐步取得支配地位;
- 强调大规模真实语料;
- 强调机器学习与知识自动获取的重要性;
- 语音识别逐步实用化;
- 开展了大规模的评测;



强调数据资源（目前）

- 各类电子词典
 - 词义词典：WordNet, HowNet, CCD
 - 语法词典：现代汉语语法信息词典
 - 语义角色：FrameNet, VerbNet
- 各种语料库
 - 英语：词性、实体、角色等，著名的 LDC
 - 汉语：分词+词性标注，汉语词义，汉语拼音 等
 - 多语言（用于机器翻译）：词对应，短语对应，句子对应，等；



强调评测

- SigHan (汉语)
- NIST (National Institute of Standard and Technology)
 - TREC
 - Open MT
 - MUC(ACE)
 - TDT
 - DUC
-



将来

- 强调多技术集成;
- 强调理性与经验方法
- 强调模块构件化和工具环境开发
- 强调知识与意义的表示与利用
- 强调实用化



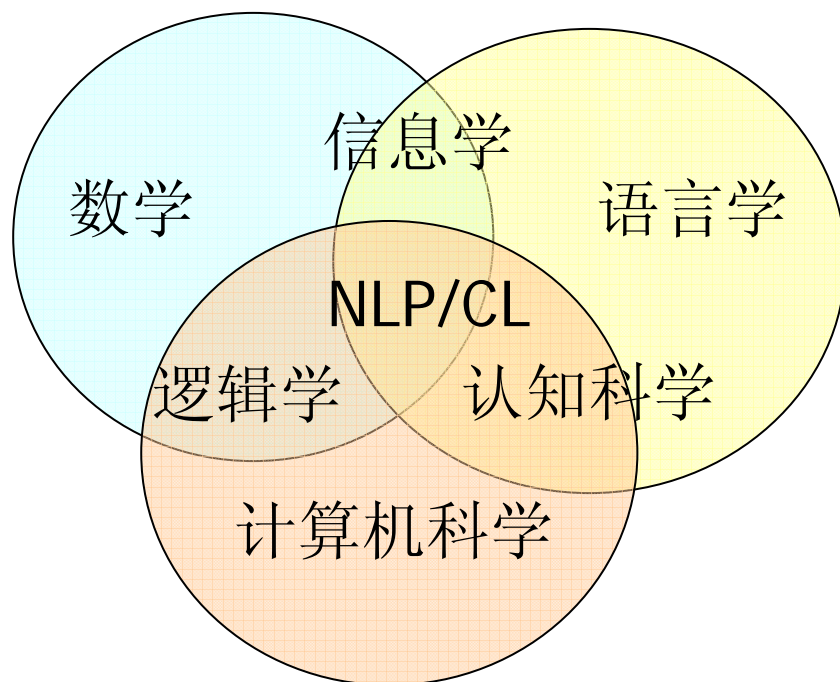
“名实之辩”

- 计算语言学（Computational Linguistics）
- 计量语言学（Quantitative Linguistics）
- 数理语言学（Mathematical Linguistics）

- 自然语言理解（Natural Language Understanding）
- 自然语言处理（Natural Language Processing）
- 人类语言技术（Human Language Technology）



小结



- 多学科交叉
- 年轻
- 应用广泛
- 挑战性（困难）



进一步阅读文献

- Hubert L. Dreyfus(1972初版, 1979修订) What Computer can't do, Harper & Row, Publishers (计算机不能做什么, 宁春岩译1986, 三联书店)
Hubert L. Dreyfus(1992) What Computers Still Can't Do: A Critique of Artificial Reason, 1992, MIT Press
- Roger Penrose (1989), The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics, Oxford University Press, 1989 (皇帝新脑)
- 袁毓林 (1993) 自然语言理解的语言学假设, 《中国社会科学》1993.1
- 宁春岩 (1985) 自然语言理解中的几个根本问题, 《语言研究》1985.2
- 詹卫东 (2000) 80年代以来汉语信息处理研究述评, 《当代语言学》2000.2



复习思考题

- 自然语言处理（计算语言学）是一个什么性质的学科？
- 人类的语言有哪些性质？
- 请举出一个算法的例子。
- 你对把“计算机”称作“电脑”有何评论？
- 人脑，电脑（机器）之间有不可逾越的鸿沟吗？