统计自然语言处理概述

刘　挺

哈工大信息检索研究室(HIT-IRLab)

2004年春

目录

- 概述
- 一个NLP的例子
- NLP的困难
- NLP方法论
- NLP的任务和瓶颈
- 统计方法示例
- 本课的主要内容

概述

NLP的概念

- 什么是自然语言处理
  - NLP, Natural Language Processing
  - 用机器处理人类语言的理论和技术
  - 区别
    - 语言处理
    - 语言信息处理（如：中文信息处理）
- 其它名称
  - 自然语言理解(Natural Language Understanding)
  - 计算语言学(CL, Computational Linguistics)
  - 人类语言技术(Human Language Technology)

- 什么是自然语言
  - 以语音为物质外壳，由词汇和语法两部分组成的符号系统。《新华词典》
  - 语言是人类交际的工具，是人类思维的载体
  - 是约定俗成的，有别于人工语言（程序设计语言）
- 什么是处理
  - 包括理解、转换、生成等

机器能够理解人的语言吗？

- 很难，但是没有证据表明不行
- 什么是理解
  - 结构主义：机器的理解机制与人相同
    - 问题在于谁也说不清自己理解语言的步骤
  - 功能主义：机器的表现与人相同
    - 图灵测试
      - 如果通过自然语言的问答，一个人无法识别和他对话的是人还是机器，那么就应该承认机器具有智能

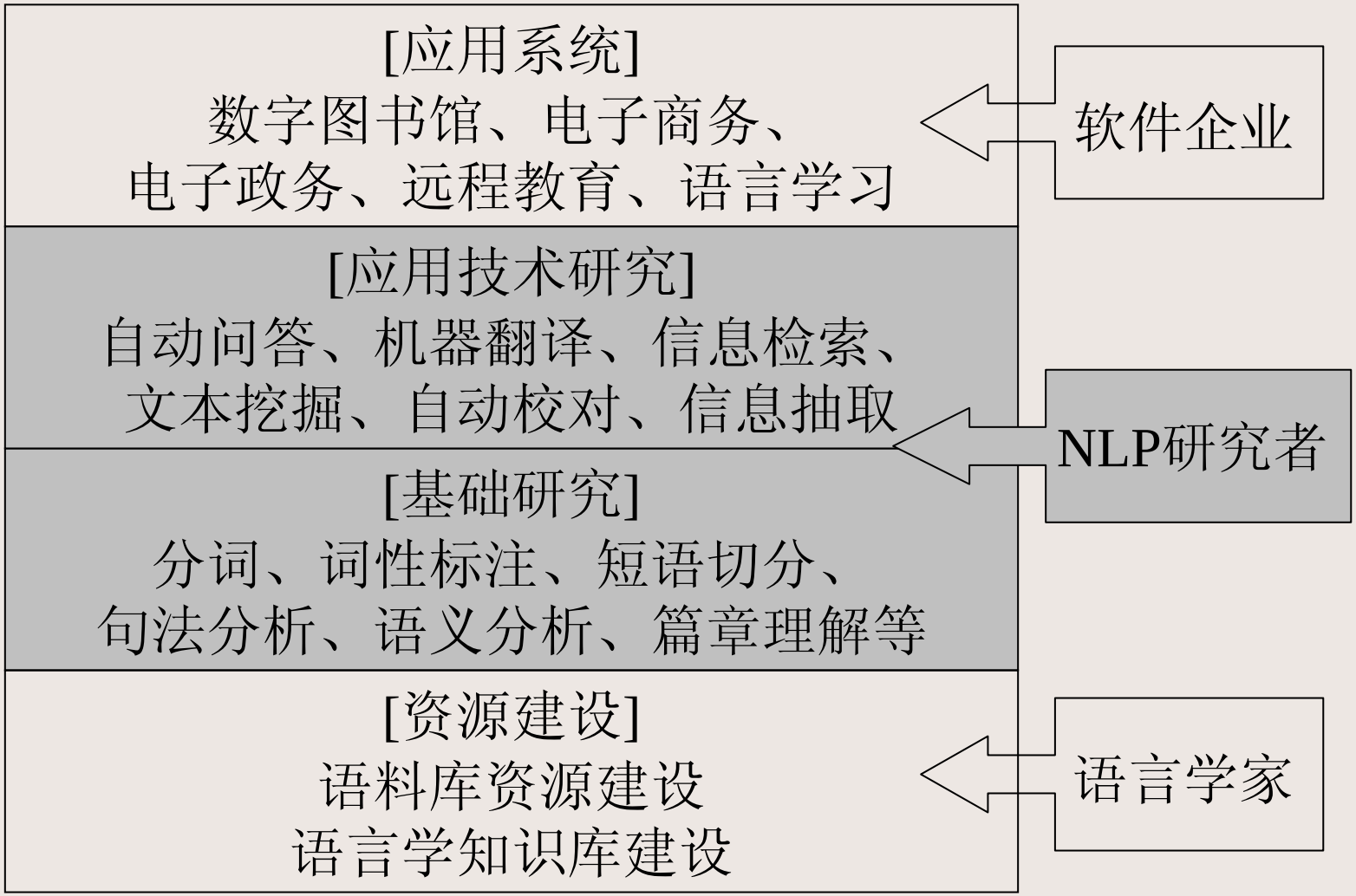
有用和能用

- NLP有用吗
  - 据统计，日常工作中80%的信息来源于语言，处理文本的需求在不断增长
  - 文本是人类知识最大的存储源，并且文本的数量在不停地增长
    - 电子邮件、新闻、网页、科技论文、 用户抱怨信
- NLP能用吗
  - 并非每一样语言处理的应用都需要深层理解
  - 中间产品陆续产生
  - 成功应用的实例
    - 微软拼音
    - 黑马中文自动校对

从智能接口到知识处理

- 智能接口
  - 功能：
    - 把现实世界中的信息送入电子世界
  - 主要成果
    - 拼音输入、手写输入、语音合成、语音输入
- 知识处理
  - 功能：
    - 对于已进入电子世界中的信息进行加工处理获得知识
  - 主要研究内容
    - 媒体的加工和管理、语言信息处理
  - 知识处理的时代已经到来！

NLP的不同层次


[应用系统]
数字图书馆、电子商务、
电子政务、远程教育、语言学习
[应用技术研究]
自动问答、机器翻译、信息检索、
文本挖掘、自动校对、信息抽取
[基础研究]
分词、词性标注、短语切分、
句法分析、语义分析、篇章理解等
[资源建设]
语料库资源建设
语言学知识库建设
软件企业
NLP研究者
语言学家

NLP的历史

- 20世纪50年代起步
  - 机器翻译、自动文摘
- 50-60年代采用模式匹配的方法
  - 60年代衰落
- 70-80年代采用面向受限域的深入理解的方法
- 90年代至今统计方法占主流
  - 随着互联网的发展而复苏
  - 互联网为NLP提供了市场需求和试验数据

NLP现状

- 仍然缺乏理论基础
- 词汇句法方面的问题尚未解决，已开始挑战语义、知识等深层课题
- 语音识别中采用的统计语言模型推动了NLP的发展，目前的统计模型在向语言深层发展
- Ontology受到普遍重视
- 开放域处理时起时落
- 一切才刚刚开始……

一个NLP的例子

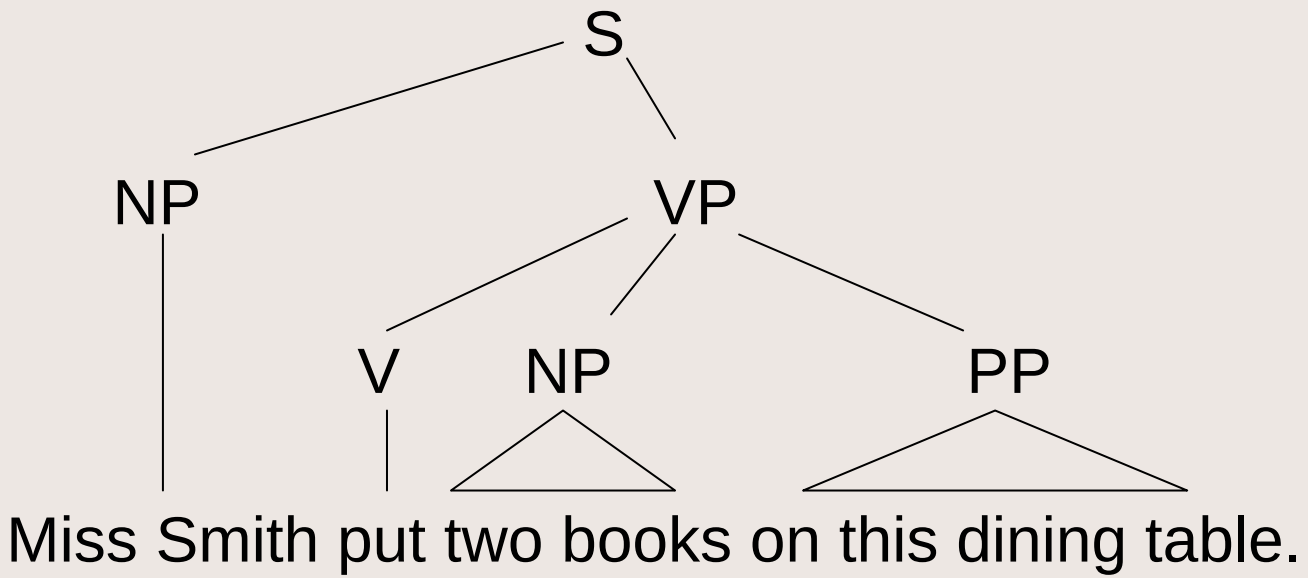
英汉机器翻译实例

- 输入英文句子：Miss Smith put two books on this dining table.

- 形态分析(Morphological Analysis)

  Miss
  Smith
  put (+ed)
  two
  book+s
  on
  this
  dining table.

- 句法分析(Syntactic Analysis)



- 词汇转换

Miss	⇒	小姐
Smith	⇒	史密斯
put (+ed)	⇒	放
two	⇒	两
book+s	⇒	书
on	⇒	在…上面
this	⇒	这
dining table.	⇒	餐桌

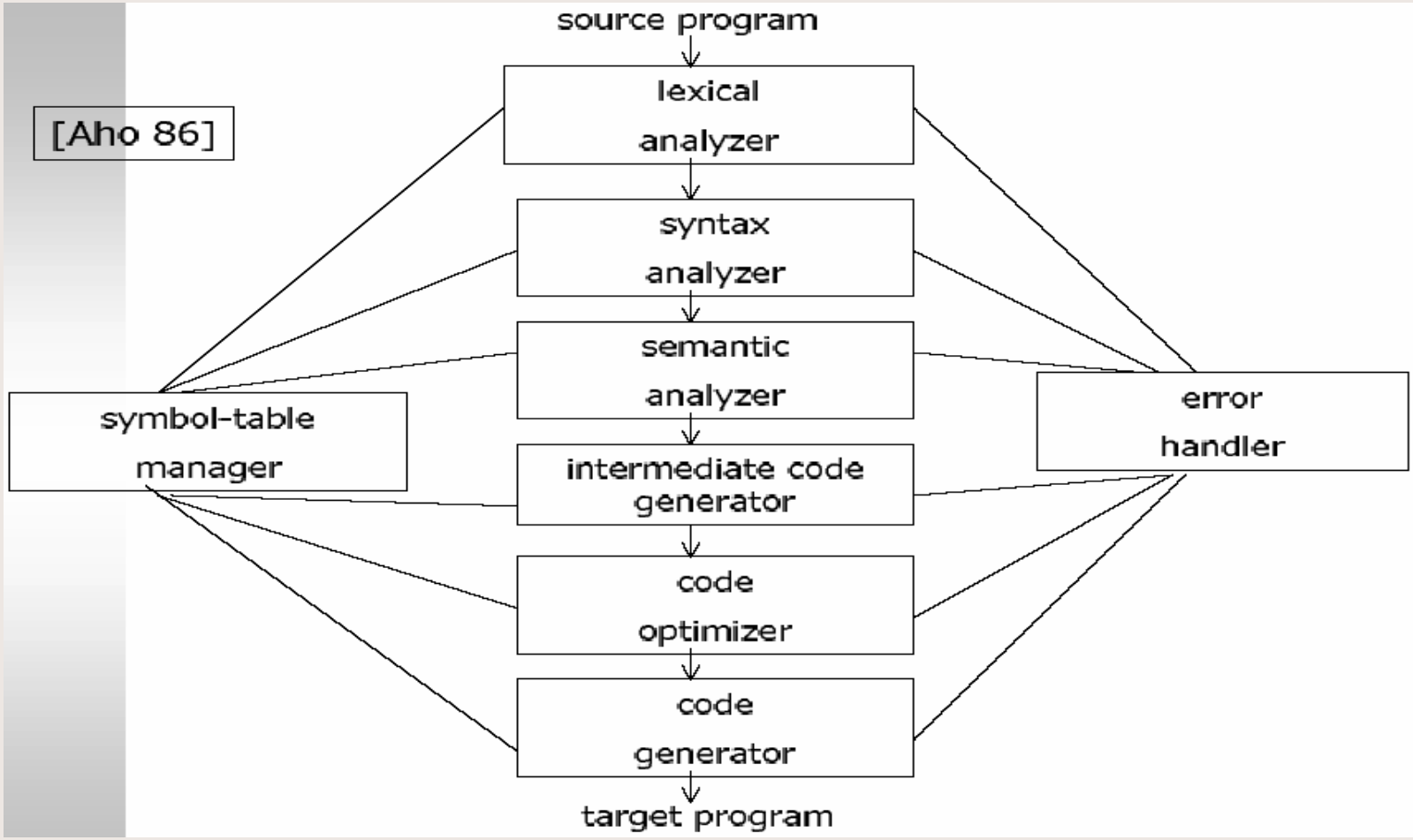
- 短语转换

小姐史密斯放两书在上面这餐桌

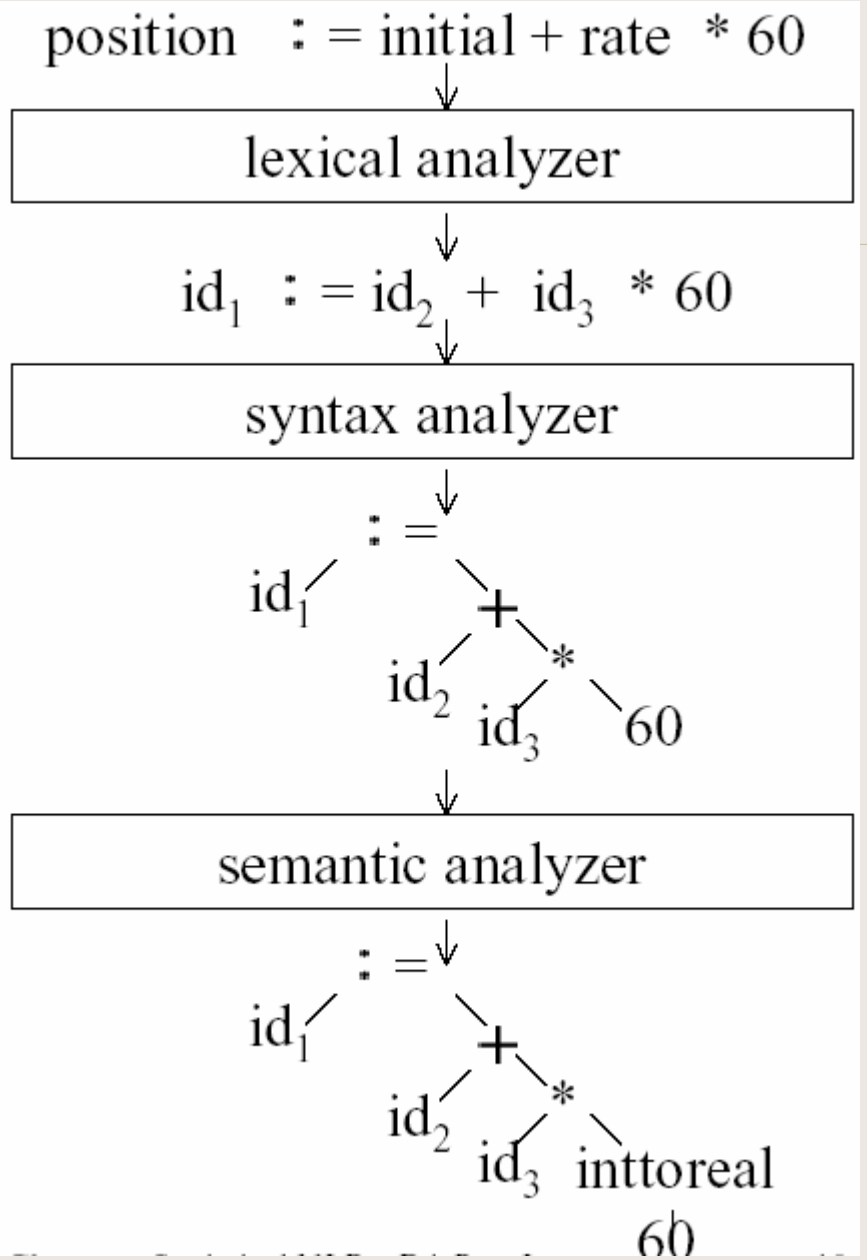
史密斯小姐放两书在这餐桌上面

- 生成
  - 史密斯小姐放两书在这餐桌上面
  - 史密斯小姐（把）两（本）书放在这（张）餐桌上面
- 最终翻译结果
  - 英文：Miss Smith put two books on this dining table.
  - 中文：史密斯小姐把两本书放在这张餐桌上面

类比编译系统


source program
lexical analyzer
syntax analyzer
semantic analyzer
intermediate code generator
code optimizer
code generator
target program
symbol-table manager
error handler
[Aho 86]

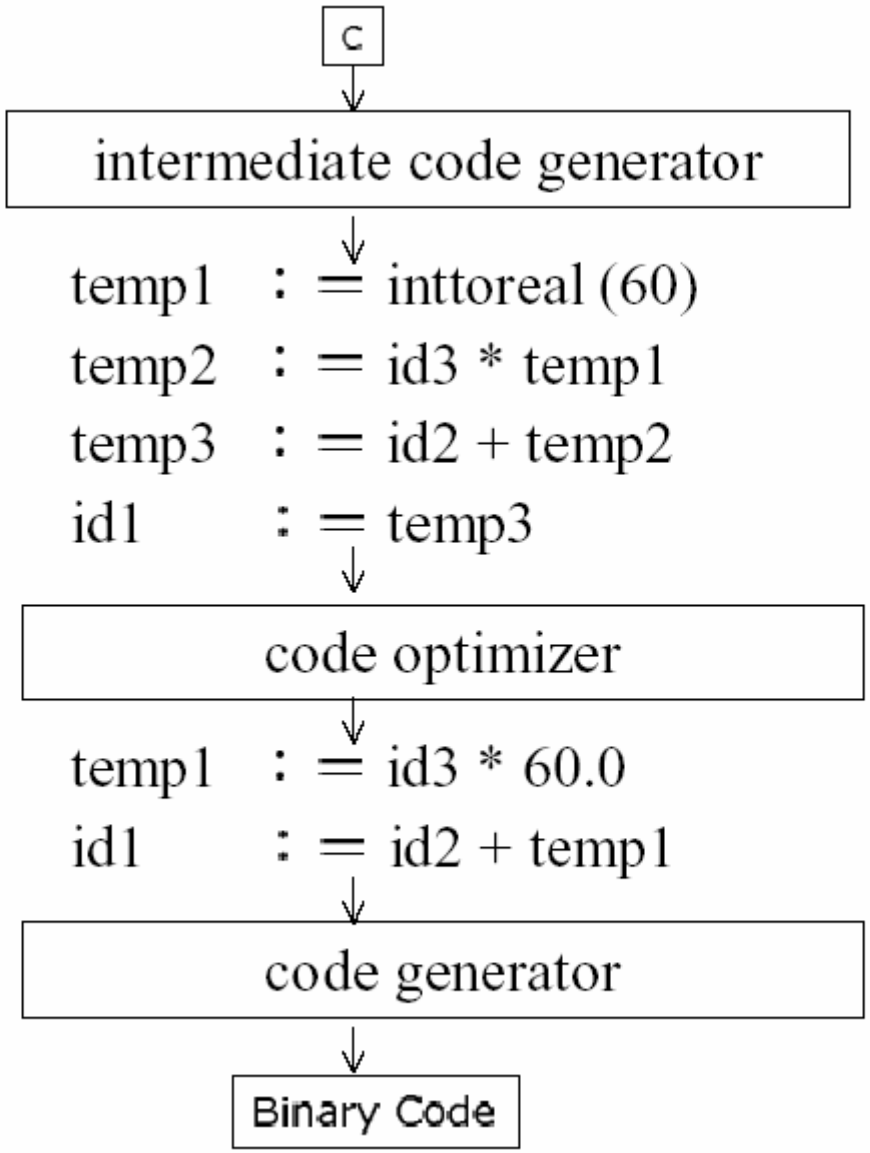
[Aho 86]



SYMBOL TABLE

1	position	…
2	initial	…
3	rate	…
4		

[Aho 86]


c
intermediate code generator
temp1 : = inttoreal (60)
temp2 : = id3 * temp1
temp3 : = id2 + temp2
id1 : = temp3
code optimizer
temp1 : = id3 * 60.0
id1 : = id2 + temp1
code generator
Binary Code

语言理解的步骤

- 文本预处理
- 句子切分
- 形态分析(Morphological Analysis)
- 分词
- 词性标注(Part-of-Speech Tagging)
- 句法分析
- 词义消歧(Word Sense Disambiguation)
- 语义关系分析
- 指代消解(Anaphora Resolution)
- 逻辑形式(Logic Form)

转换与生成

- 处理
  - 翻译
    - 运用翻译规则或统计模型等，将源语言的内部表示转换为目标语言的内部表示
  - 文摘
    - 对源语言文本进行压缩，提取出关键句子
- 生成(Generation)
  - 模拟人类写作的过程，生成符合逻辑的连贯的文本

NLP的困难

歧义(Ambiguity)

病构(Ill-Formedness)

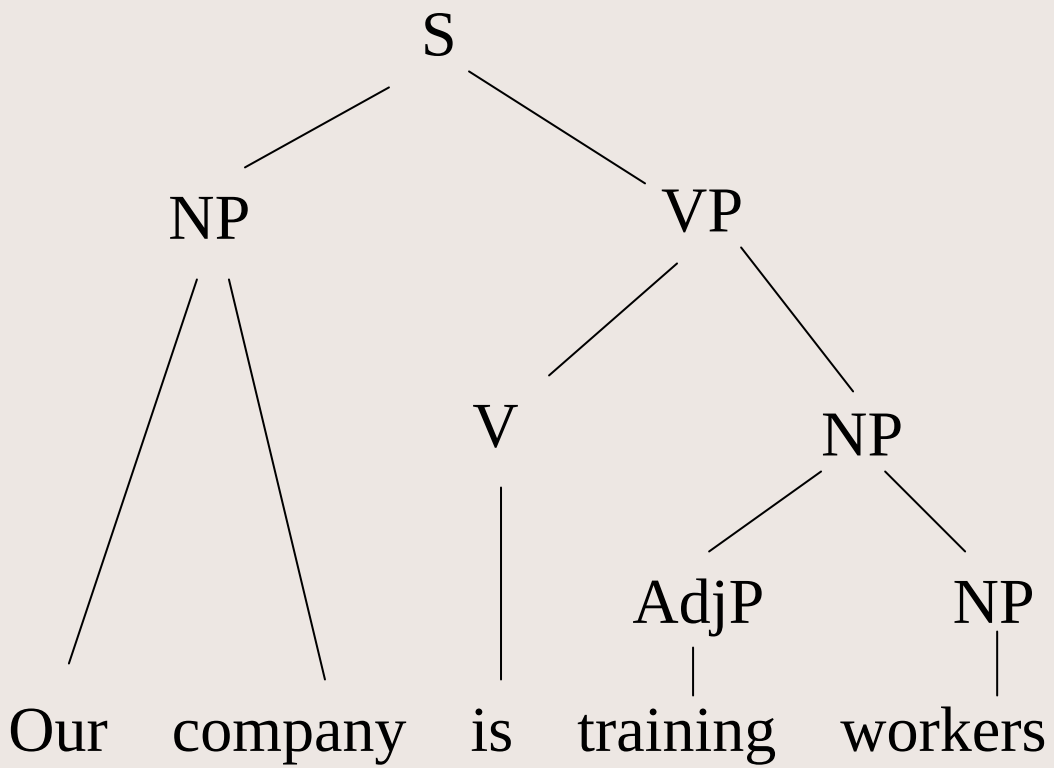
——台湾：苏克毅

歧义

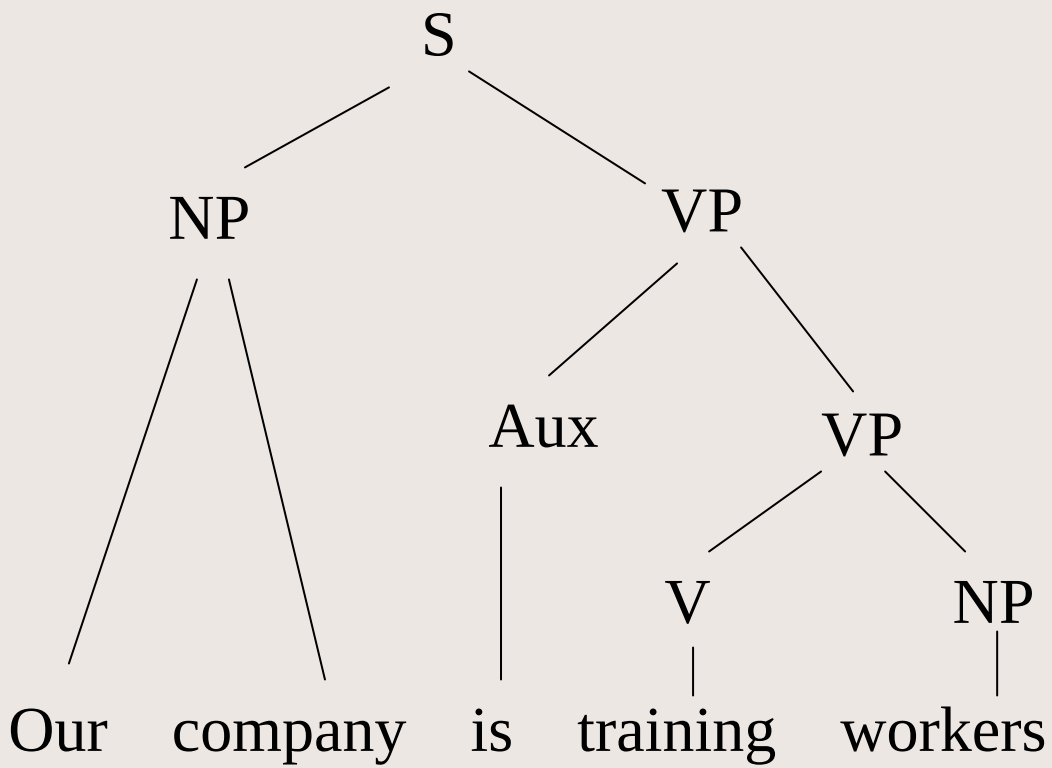
- 注音歧义
  - 快乐(le4)的单身汉
  - 火红的第五乐(yue4)章
- 分词歧义
  - 交集歧义
    - 研究/ 生命/ 的/ 起源
    - 研究生/ 命/ 的/ 起源
  - 组合型歧义
    - 他/ 从/ 马/ 上/ 下来
    - 他/ 从/ 马上/ 下来

- 分词歧义
  - 和未登录词绞在一起
    - 刘挺/ 拔/ 出/ 宝剑
    - 刘/ 挺拔/ 出/ 宝剑
  - 多交集字段的歧义
    - [结 合][成 分][子 时]
  - 有的歧义无法在句子内部解决
    - 乒乓球拍卖完了
- 短语歧义
  - [咬死猎人]的狗
  - 咬死[猎人的狗]

Our company is training workers(1)



Our company is training workers(2)



- 词义歧义
  - 打[玩]乒乓球
  - 打[编制]毛衣
  - 打[通讯]电话
  - ……
- 语用歧义
  - “你真讨厌！”

病构

- 真实文本的语言现象非常复杂，不规范，不干净
  - 未登录词(Unknown Words)
  - 已知词的新用法
    - 例子：Please xerox a copy to me.
  - 不合乎语法的句子
    - 例子：他非常男人。（名词不能受程度副词修饰）
  - 不合乎语义约束的搭配
    - 例子：My car drinks gasoline like water.
  - 由于作者疏忽造成的错误
  - 真实的语言是非常脏的

重述(Paraphrasing)

- 举例
  - 毛泽东出生于1893年
  - 毛泽东出生在1893年
  - 毛泽东诞生于1893年
  - 毛泽东同志是1893年出生的
  - 毛主席生于1893年
  - 毛泽东生于光绪6年（虚拟的）

层间循环依赖问题

- 循环依赖
  - 高层模块建立在底层模块分析的基础上
  - 底层模块需要高层模块的指导才能准确分析
- 如何克服这种致命的矛盾
  - 简单级联
    - 每层的准确率是90%，如果系统分6层，最终结果是：53%；即使每层95%，最终结果73%
  - 一体化：如分词词性标注一体化
  - 反馈
  - 黑板结构
- 人是怎么做的？
  - 人在瞬间综合运用各个层面的知识

NLP方法论

语言处理的科学内容

- 语言学的任务
  - 刻画和解释语言现象
    - 人类是如何获取和理解语言的
    - 理解语言和世界的关系
    - 理解语言在通讯时的结构
- 人们在说些什么
  - 覆盖语言结构的各个方面
- 人们说的事情和世界怎样联系在一起

语言的三个层面

- 句法(Syntax)
  - 形式结构：主语、谓语、宾语等
  - 用形态变化(-ing, -ed)、虚词（“的”、“了”）和词序等体现
- 语义(Semantics)
  - 语言和世界的映射关系
  - 施事、受事、工具等
- 语用(Pragmatics)
  - 语言交际目的
  - 主题、述体、焦点

句法、语义、语用的区别

- 句法结构相同，语义不同
  - “吃苹果”，“吃食堂”
    - 句法：动宾结构
    - 语义分别为：动作-对象关系，动作-地点关系
- 语义相同，句法结构不同
  - “吃了苹果”，”苹果吃了”
    - 语义：动作-对象
    - 句法分别为：动宾关系和主谓关系
- 语用
  - 语义相同，语用有别
    - 主席台上摆着鲜花（主席台是旧信息，鲜花是新信息）
    - 鲜花摆在主席台上（主席台是旧信息，鲜花是新信息）

理性主义和经验主义

- 理性主义者(Rationalist)
  - 1960-1985: 理性主义是主流
  - 他们的信念
    - 乔姆斯基
    - 先天语言能力
- 对于语法的描述
  - 形成基于规则的传统语言处理技术
- 句法规则的确抓住了语言的主要模式
- 什么是语言中最普遍的模式呢，是否需要量化？

理性主义的问题

- 语言的变化是渐变的
  - 比如：“打”电话，究竟从那一天开始“打”被赋予了通讯的意义呢
- 基于规则的方法需要大量的人工操作，人类总结的规则不完备、不一致，规则多了相互冲突，难以对抗复杂的语言现象

经验主义者

- 信念
  - 孩子的大脑只能做一些普通的操作：连接、模式识别、一般化。孩子从丰富的信号输入中学习到了语言的结构
- 设定一个语言模型，推导出参数值
  - 形成今天的基于统计的语言处理技术
  - 对每一种语言现象均给出统计量化指标
- 意义：“观其伴，知其意”

经验主义

- 我们生活在一个充满不确定和不完整信息的世界里
- 人类的认知是一个随机现象
- 语言也是一个随机现象
- 对没有见过的语言现象进行估计
- 复杂的概率模型

理性主义和经验主义的差别

- 它们描述了不同的事情
- 理性主义试图去描写人脑中的模型
  - 结构主义者
- 经验主义试图去描写实际出现的语言
  - 功能主义者
- 外部语言是内部语言的非直接的事实

进一步探讨

- 从九十年代初期开始，统计方法开始成为自然语言处理的主流
- 规范的语言和非规范的语言之间没有明确的界限
- 统计还是非统计，界限也比较模糊
- 追求纯净，还是实用
- 自然语言处理尚不存在统一的数学基础
  - 概率模型、信息论和线性代数

语言工程

- 近来，人们更有兴趣解决工程实际问题
- 人们处理真实世界中的语料，并客观地比较不同方法的优劣
- 面向真实文本的评测，是科学研究和技术开发进一步统一起来。
  - 90年初的汉语分词系统仍未考虑“未登录词”问题，那时已经宣称分词结果达到90%以上，其实只是解决了部分歧义问题。90年代中后期才开始面向真实文本的处理。

NLP的任务和瓶颈

NLP的性质

- NLP需要的知识非常复杂
- 理解语言的过程是动态的，不是静态的
- NLP需要的知识大多是归纳的，不是演绎的
- 人也不一定能够出一致的理解结果
  - 存在Upper Bound（上限）
- NLP是一个非确定性过程
- 对歧义的限制和系统的覆盖率矛盾
- 领域词典不充分

NLP系统的主要任务

- 知识表示
  - 产生式
  - 谓词逻辑
  - 语义网络
  - 概念从属理论（CD理论）
- 知识控制策略
  - 知识的冲突

NLP系统的主要任务

- 知识集成
  - 从多个知识源获取的不同层面，不同性质的只是如何融合在一起
- 知识获取
  - 恳谈式
  - 内省式
  - 机器学习

NLP的瓶颈

- 知识获取(Knowledge Acquisition)
  - 知识获取和知识表示相关联
    - 规则：人工知识
    - 参数：适合机器学习
  - 混合方法(Hybrid Approach)
    - 人设计模型
    - 机器训练参数

统计方法示例

从语料库中学习

- 语料库
  - Corpus, Corpora
  - 文本的集合
    - 可以原始的文本（生语料库）
    - 也可以是带标记的文本（熟语料库）
- 语料库是统计NLP的知识来源

搭配知识

- 在一个窗口中抽取的搭配知识可以应摄影深层的语义关系
- 例子：
  - 维护国家的利益
    - VP(V+N)+de+N or V+NP(N+de+N)?
  - 在语料库中有:
    - 维护我们的利益, 维护中国的利益, ……
    - 国家利益不容侵犯, 损害国家的利益, ……

语料库资源

- Brown Corpus
  - 带词性标记，一百万词
  - 布朗大学
  - 平衡语料库
  - 美国英语
  - 1960s-1970s
- Lancaster-Oslo-Bergen (LOB)
  - British English of the Brown corpus
- Susanne corpus
  - Brown语料库的子集，13万词

Lexical Resources(II)

- Penn Treebank（宾州树库）
  - 美国宾西法尼亚大学开发
  - 取材华尔街日报
  - 以开发中文树库，但规模有限
- Canadian Hansards
  - 加拿大议会双语文本
- WordNet
  - 语义词典，免费使用

- HowNet
  - 中文语义词典.
- 北京大学语法词典
- 北大-富士通《人民日报语料库》
  - 半年的《人民日报》
  - 带词性标注

举例

- 一篇短篇小说
  - 作者：Mark Twain
  - 小说名：Tom Sawyer
  - 词数(Word tokens)
    - 71,370
  - 词形数(Word types)
    - different things present
    - 8,018
  - 平均每个词形出现: 8.9次

最高频率的词汇

- English
  - the 3332
  - and 2972
  - a 1775
  - to 1725
  - of 1440
- Chinese
  - 的 5%

一些结果

- 词频　　具有该词频的词的数目

1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
>100	102

困难

- 一些结果
  - 最高频的100个词覆盖了全部词汇出现次数的一半
  - 一半的词汇在语料库中只出现一次
  - 90%的词形出现10次或更少
  - 文本中的12% 是出现3次或者更少的词
- 很难预测那些很少出现或者干脆在语料库中从未出现的词的行为

齐普夫定律

- 讲者和听者就试图使用最小的力气
  - 讲者希望：使用最少的词汇，没有标点空格
  - 听者希望：使用较多的词汇，丰富的标记
- 什么是齐普夫定律？
  - 在一个大的语料库中统计词频，然后将词按照词频从高到低的顺序排列成一张表
  - 一个词的词频 f 和它在表中的序号r之间存在如下关系：

    $f \propto 1/r$ or $f \cdot r = k$, k是一个常数

数据

Word	Freq	Rank	f·r	Word	Freq	Rank	f·r
the	3332	1	3332	and	2972	2	5944
a	1755	3	5235	he	877	10	8770
but	410	20	8400	be	294	30	8820
there	222	40	8880	one	172	50	8600
two	104	100	10400	turned	51	200	10200
you’ll	30	300	9000	name	21	400	8400
comes	16	500	8000	family	8	1000	8000
brushed	4	2000	8000	sins	2	3000	6000
Could	2	4000	8000	Applausive	1	8000	8000

词频的分布

- 齐普夫定律是对人类语言词频分布的一个粗糙而有用的描述:
  - 非常常用的词很少
  - 中频词的数量中等
  - 大量低频词
- 从语料库中，我们能够观察到少数高频词的丰富的信息，而对大量低频词却观察不到足够数量的信息

词义和词频的关系

一个词的词义的数量和该此词频的平方根成正比关系

$$m \propto \sqrt{f} \quad \text{或} \quad m \propto \frac{1}{\sqrt{r}}$$

频率等级： 10,000(sqrt=100)

average about 2.1 meanings, 210

频率等级： 5000(sqrt=70.7)

average about 3 meanings, 212.1

频率等级： 2000(sqrt=44.7)

average about 4.6 meanings, 205.62

词频和词长

- 词频和词长是反比例关系
- 短词经常被使用
  - “in”, “of”, ……
  - “的”，“了”
- 这符合通讯编码理论

搭配

- 搭配(Collocations)
  - 复合词(*disk drive*)
  - 短语动词(*make up*)
  - 其它固定短语 (bacon and eggs).

二元搭配

- 例子：
  - 80871 of the
  - 58841 in the
  - 26430 to the
  - 21842 on the
  - 21839 for the
  - 18568 and the
  - ……

对搭配进行过滤

- 根据词性过滤掉一些搭配，例如虚词，从而获得真正有意义的搭配
- 最高频的搭配模式是：
  - 动词+名词
  - 形容词+名词
  - 名词+名词

有意义的搭配

- 例子：
  - 11487 New York A N
  - 7261 United States A N
  - 5412 Los Angeles N N
  - 3301 last year A N
  - 3191 Saudi Arabia N N
  - 2699 last week A N
  - 2514 vice president A N

KWIC

- Key Word In Context(KWIC)
  - The librarian “showed off” - running hither and
  - lady teachers “showed off” - bending sweetly
  - his lip and showed the vacancy
  - pen. Then he showed Huckleberry how to
  - and his eyes showed the fear that was upon
  - ……

KWIC的使用

- 对语言学家有用
  - 北京语言大学的宋柔教授，开发了基于PATTree的为语言学家服务的词汇用法检索系统
- 有利于统计句法分析器等的开发

本课的主要内容

- 预备知识
  - 导论
  - 数学基础
  - 语言学基础
  - 语料库
- 词汇
  - 搭配
  - 统计推理
  - 词义消歧
  - 词汇知识获取

- 语法
  - 马尔科夫模型
  - 词性标注
  - 概率上下文无关文法
  - 概率句法分析器
- 应用
  - 机器翻译
  - 聚类和分类

参考书

- Manning, C.D.
  - Foundations of Statistical Natural Language Processing, MIT Press, 1999
- Jurafsky, D.
  - Speech and Language Processing, Prentice Hall, 2000
- Allen, J.
  - Natural Language Understanding, The Banjamins / Cummins Publishing Co. 1994

NLP领域的学术会议

- 主要国际会议
  - ACL(2004：巴塞罗那）
    - Association of Computational Linguistics
  - Coling (2004：维也纳)
  - IJCNLP（2004：海南）
  - EACL(European Chapter of ACL)
  - ANLP(Applied NLP)
  - SIGIR(SIG Information Retrieval)
  - TREC(Text REtrieval Conference)
- 主要国内会议
  - JSCL(全国计算语言学联合学术会议)

NLP领域主要学术机构

- 国外
  - 美国
    - CMU-LTI(Language Technology Institute)
    - 南加州大学ISI(Information Science Institute)
    - 宾西法尼亚大学
  - 日本
    - ATR
  - 加拿大阿尔博塔：林德康
- 国内外企和港台
  - 微软研究院：周明、李航、高剑峰
  - 香港城市大学：黄锦辉
  - 台湾：苏克毅、陈克俭、简立峰

国内研究机构

- 北京
  - 清华：孙茂松、周强
  - 北大：俞士汶、孙斌
  - 中科院自动化所：徐波、赵军
  - 声学所：黄曾阳
  - 北京语言大学：宋柔、荀恩东
- 京外
  - 复旦：吴立德、黄萱菁
  - 交大：王永成
  - 东北大学：姚天顺、朱靖波
  - 厦门大学：史晓东

谢谢！