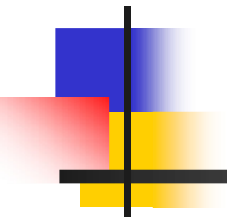


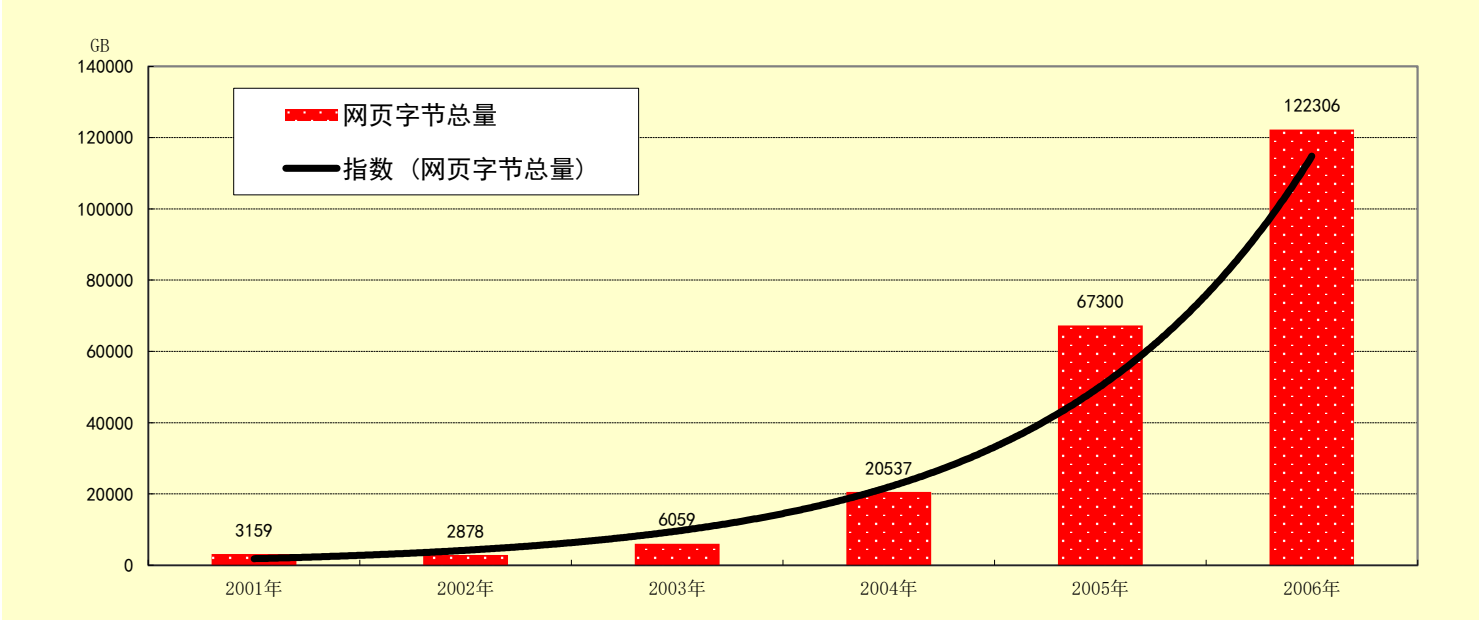
第1章 绪论

1.1 基本概念

语言是思维的载体，是人际交流的工具。

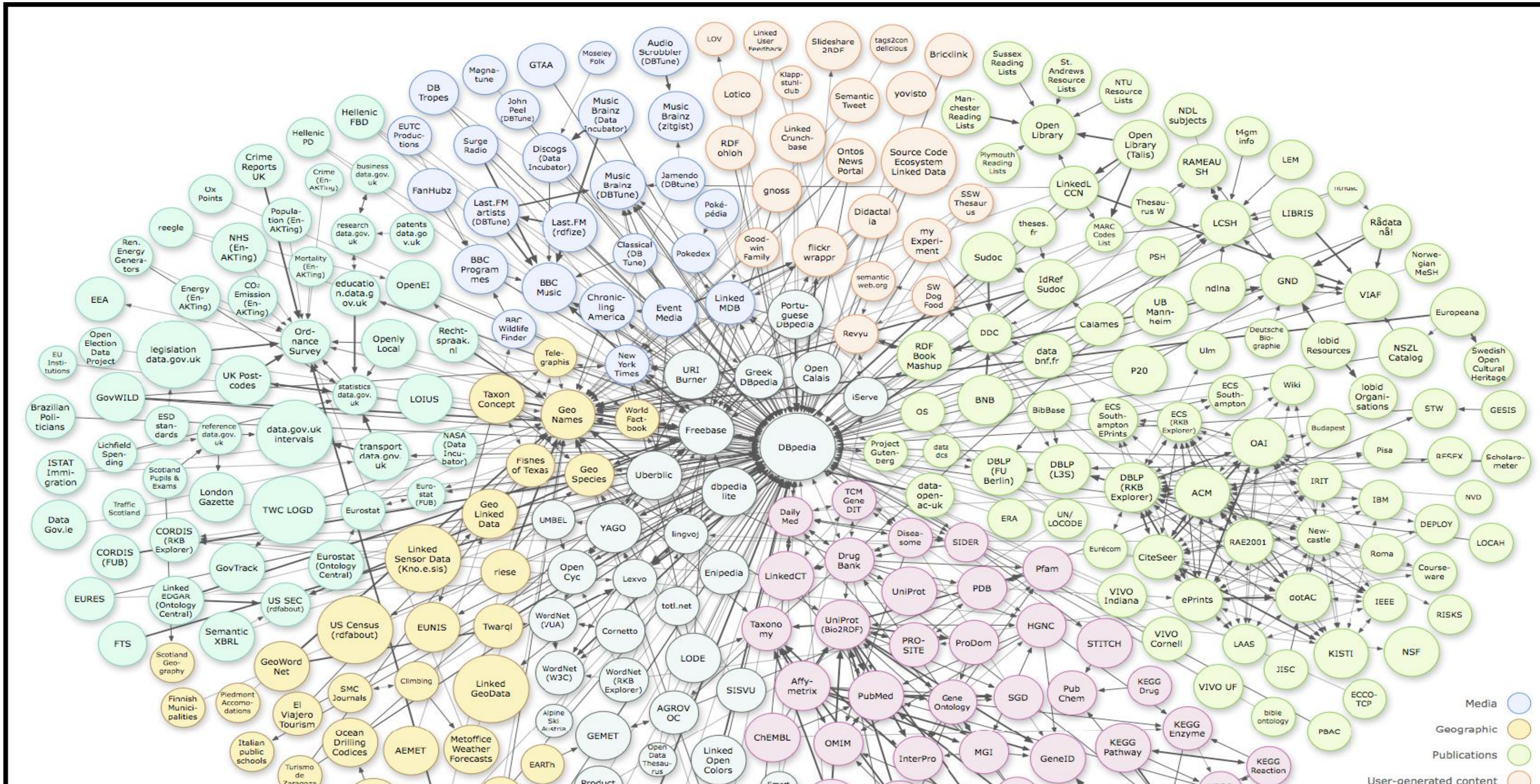
语言的两种属性—文字和声音。

无处不在的网络、通讯和堆积如山的文档，构成了当今社会信息爆炸的基本特征，也使人们面临许多难以克服的困难和障碍。



- 中文网页检索的准确率不高（40~50%？）

1.1 基本概念



人工难以应对的数据“灾难”！

1.1 基本概念

- 如何让计算机能够理解自然语言文本，懂得人的意图和心声？

- 一如何让计算机实现海量语言文本的自动处理、挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务？

1.1 基本概念

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

– 将语言做为计算对象,并研究相应的算法、模型

– 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

1.1 基本概念

近几年来，自然语言处理研究得到了前所未有的重视和长足的进展，并逐渐发展成为一门相对独立的学科而倍受关注。

自然语言处理技术不断与语音识别(speech recognition)、语音合成(speech synthesis)等语音技术相互渗透和结合形成新的研究分支。

1.2 自然语言处理研究的内容

按照应用目标划分

1、机器翻译 **(Machine translation, MT)**：实现一种语言到另一种语言的自动翻译。

应用：文献翻译、网页翻译和辅助浏览等。

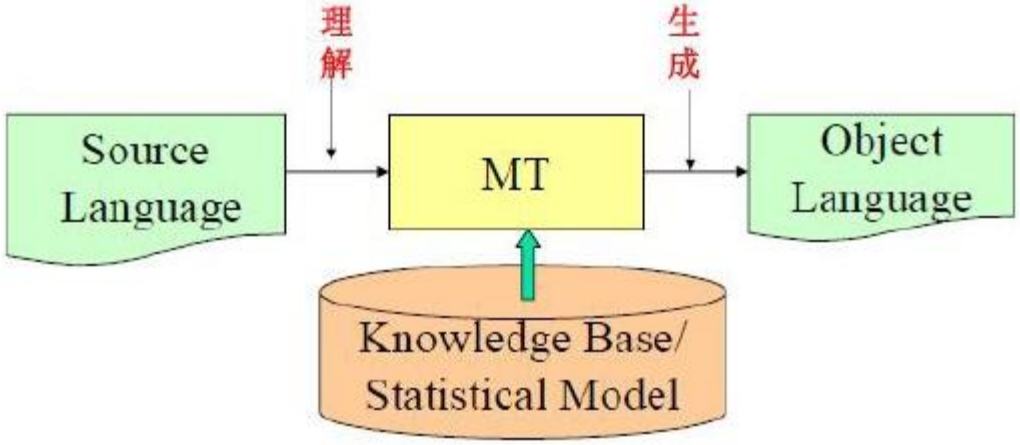
实用系统：**Systran (http://www.systransoft.com)**

36种语言对，20个专门领域。

1.2 自然语言处理研究的内容

两个过程：

- 原语言的分析和理解
- 目标语言的生成



1.2 自然语言处理研究的内容

机器翻译现状和对机器翻译的认识

例1: The spirit is willing, but the flesh is weak.

精神是愿意的, 但骨肉是微弱的。(Systran)
（心有余，而力不足。）

例2: Out of sight, out of mind.

出于视域, 在头脑之外。 (Systran)
（眼不见，心不烦。）

1.2 自然语言处理研究的内容

2、信息检索 (Information retrieval)：

利用计算机系统从大量文档中找到符合用户需要的相关信息。

面向多语言的信息检索叫做跨语言信息检索(Cross-language information retrieval)。

代表系统：Google: http://www.google.com

百度：http://www.baidu.com.cn/

1.2 自然语言处理研究的内容

信息检索难点

新闻 **网页** 贴吧 知道 MP3 图片 视频 地图

和服 | 百度一下

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。

1.2 自然语言处理研究的内容

3、自动文摘（Automatic summarization）

将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。

中国空军航空兵赴南海常态化战斗巡航

2016年07月19日09:10　来源：新华社　分享到：

中国空军新闻发言人申进科大校7月18日在北京宣布：中国空军近日组织了航空兵赴南海战斗巡航。这次南海战巡，空军出动轰-6K飞机赴黄岩岛等岛礁附近空域进行了巡航。

申进科介绍，中国空军航空兵此次赴南海例行性战斗巡航，紧贴使命任务和实战准备，轰-6K和歼击机、侦察机、空中加油机等遂行战巡任务，以空中侦察、对抗空战和岛礁巡航为主要样式组织行动，达成了战斗巡航目的。

申进科表示，中国空军航空兵赴南海战斗巡航，旨在推动海上方向实战化训练深入发展，提升应对各种安全威胁的实战能力，维护国家主权和安全。他表示："根据有效履行空军使命任务的需要，空军航空兵赴南海战斗巡航，将继续常态化进行。"

空军新闻发言人指出，南海诸岛自古以来就是中国领土，中国在南海的主权和权益不容侵犯。中国空军坚定不移捍卫国家主权、安全和海洋权益，坚决维护地区和平稳定，应对各种威胁挑战。（记者张玉清、张汨汨）

中国空军近期组织了航空兵对南海进行了常规巡逻。空军新闻发言人指出中国在南海的主权和利益不容侵犯。外国专家认为不确定性是该动作如何影响亚太与国际社会舆论。

1.2 自然语言处理研究的内容

4、文档分类 (Document categorization)

目的是利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。

应用：图书管理、内容管理、信息监控等

1.2 自然语言处理研究的内容

5、问答系统 (Question-answering system)：

通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。

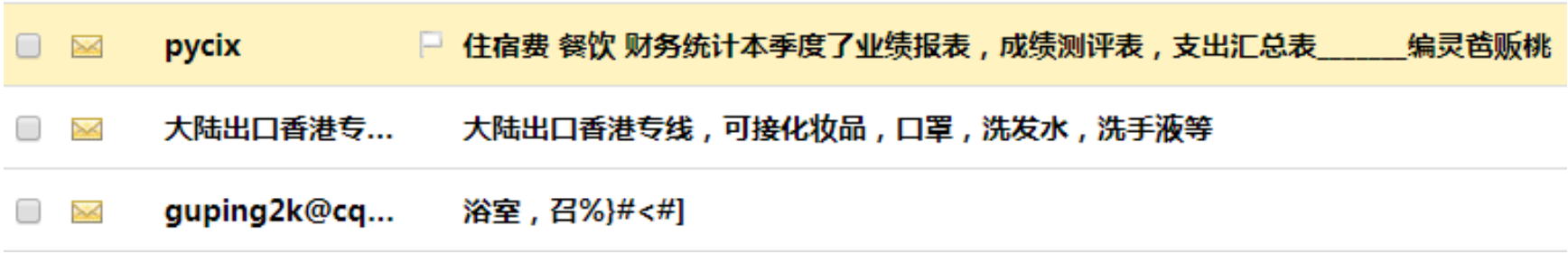
应用：人机对话系统、客户服务系统等

颐和园怎么走？
香港明天天气如何？
问航班/火车时刻
网上找人
网上购物问价格

1.2 自然语言处理研究的内容

6、信息过滤 (Information filtering)：通过计算机系统自动识别和过滤那些满足特定条件的文档信息。

应用：网络有害信息过滤、信息安全等



1.2 自然语言处理研究的内容

7、语言教学 (Language teaching)：

借助计算机辅助教学工具，进行语言教学、操练和辅导等。

应用：语言学习等

MyAccess!是一款美国的辅助写作评价工具，可以为学生提供一个写作环境，学生可以得到结构化的反馈和诊断报告，可以根据这些结果修改自己的作文，改进写作技巧。

1.2 自然语言处理研究的内容

8、文字识别 (Character recognition)：

通过计算机系统对印刷体或手写体等进行自动识别，将其转换成计算机可以处理的电子文本。

应用：文字输入、识别等

9、文字编辑和自动校对 ：

对文字拼写、用词、甚至语法、文档格式等进行自动检查、校对和编排。

应用：排版、印刷和书籍编撰等

拼写校对：我们要京城（精诚）合作

1.2 自然语言处理研究的内容

10、语音识别(speech recognition)：

将输入计算机的语音信号转换成书面语表示。

应用：文字录入、人机通讯、语音翻译等等。

困难：大量存在的同音词、近音词、集外词、口音等等。

例如：输入：美欧贸易摩擦升级

识别结果：美欧贸易摩擦生机

11、文语转换 (text-to-speech)：

将书面文本自动转换成对应的语音。

1.2 自然语言处理研究的内容

12、情感及观点分析：

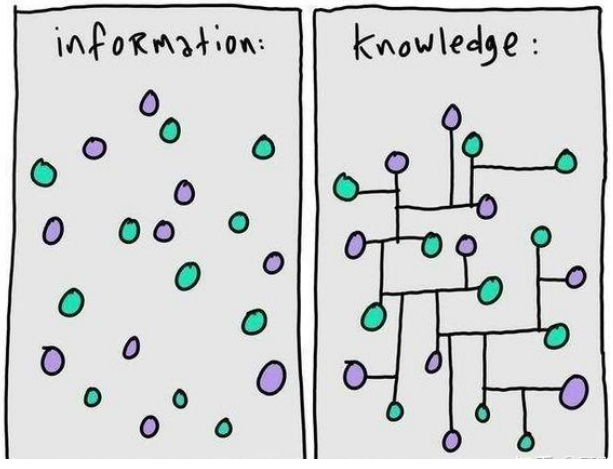
大量应用需要情感与观点分析。
如商品评论，服务质量，影评等。

- 情感与观点分析要做什么？
  - 观点是什么？带有怎样的情感色彩（正面/负面）？
  - 谁发表的观点或表达的情感？
  - 针对的问题及对象是什么？

1.2 自然语言处理研究的内容

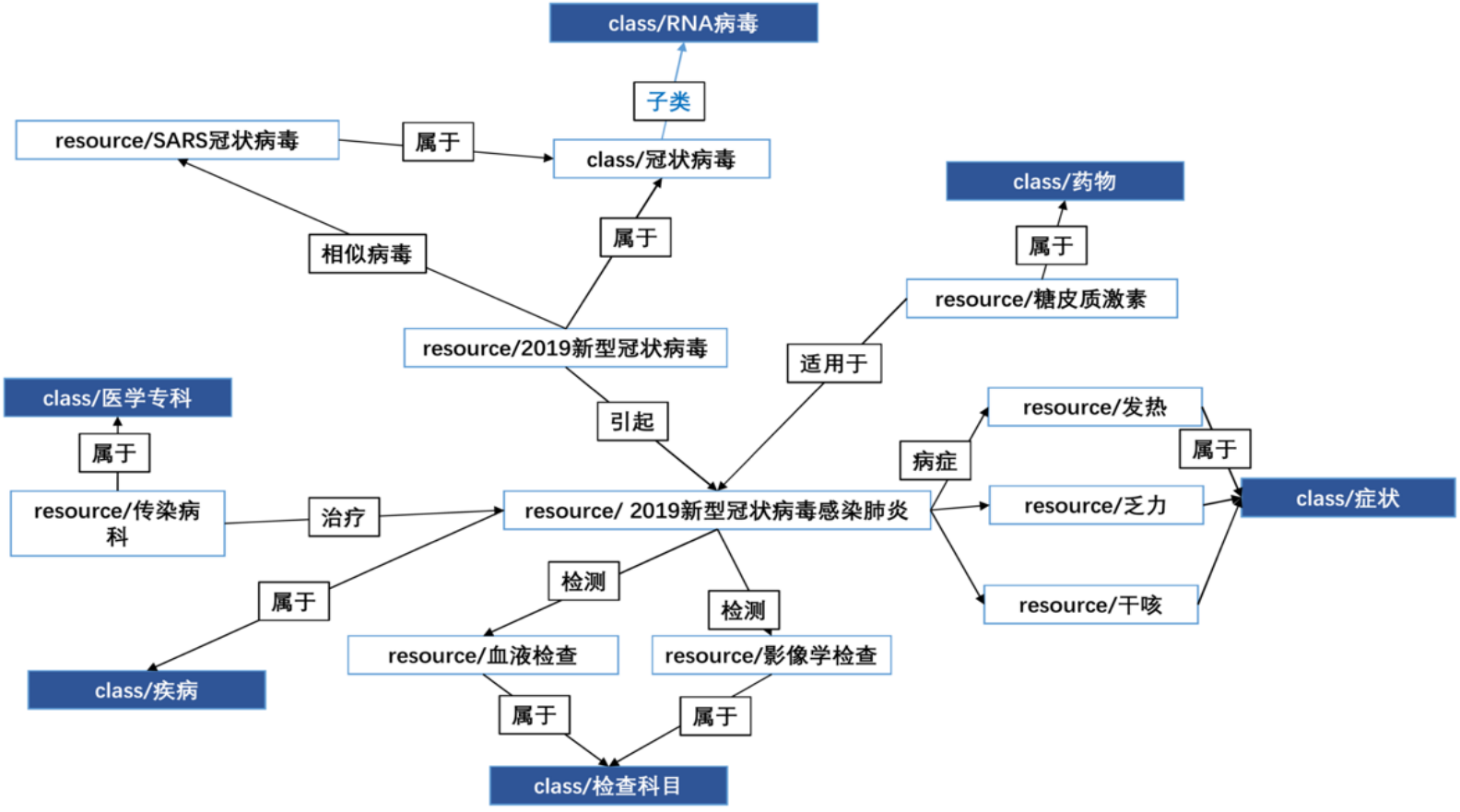
13、知识图谱：

知识图谱是一种揭示实体之间关系的语义网络。



1.2 自然语言处理研究的内容

13、知识图谱



1.2 自然语言处理研究的内容

14、文本自动生成

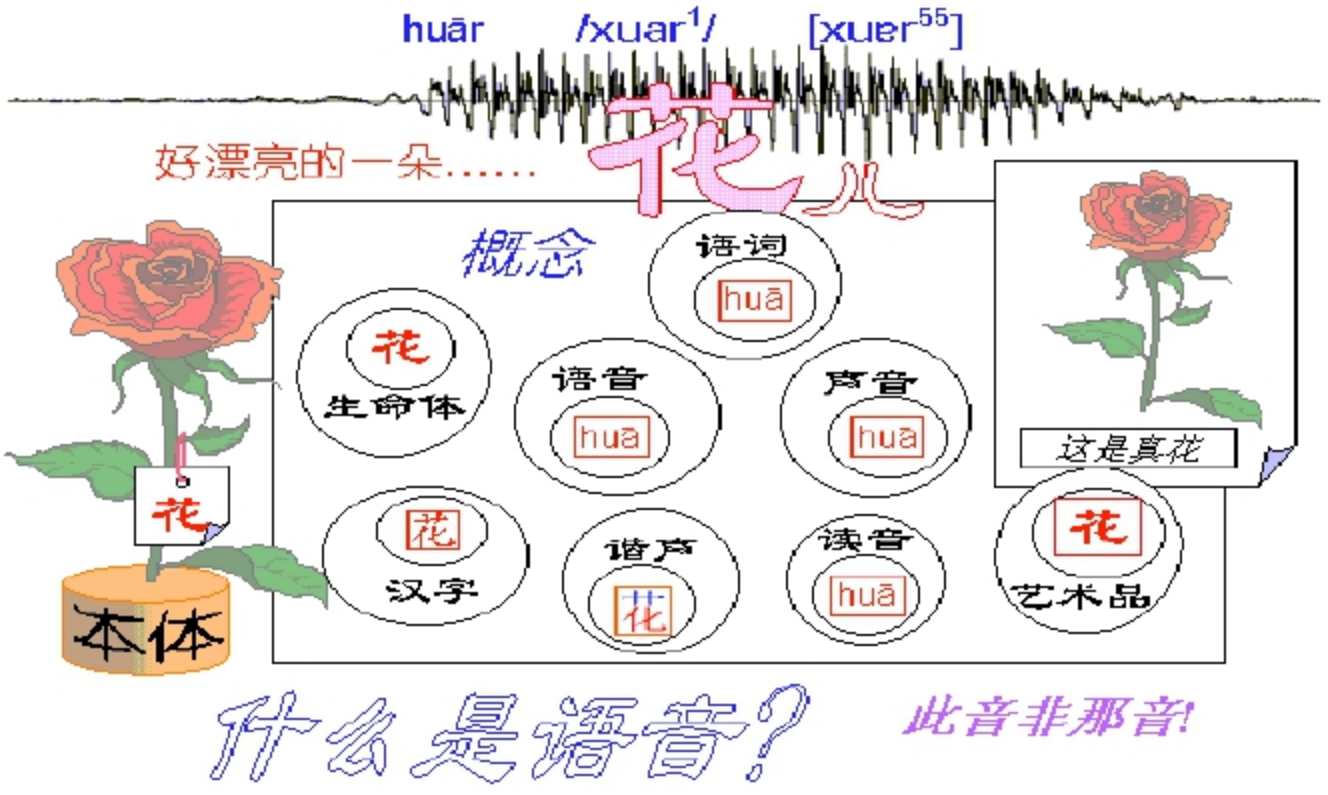
文本自动生成旨在实现机器像人一样写作,减少语言工作人员的工作量

应用：新闻稿撰写、自动作诗。

http://couplet.msra.cn/app/couplet.aspx

1.3 自然语言处理的基本问题

语音学(Phonetics) 问题：研究词及其语音的关联。



1.3 自然语言处理的基本问题

形态学 **(Morphology)** 问题：研究词是如何由词素构成的，即研究词的内部结构。

词素（morphemes）→ 词 （word）？

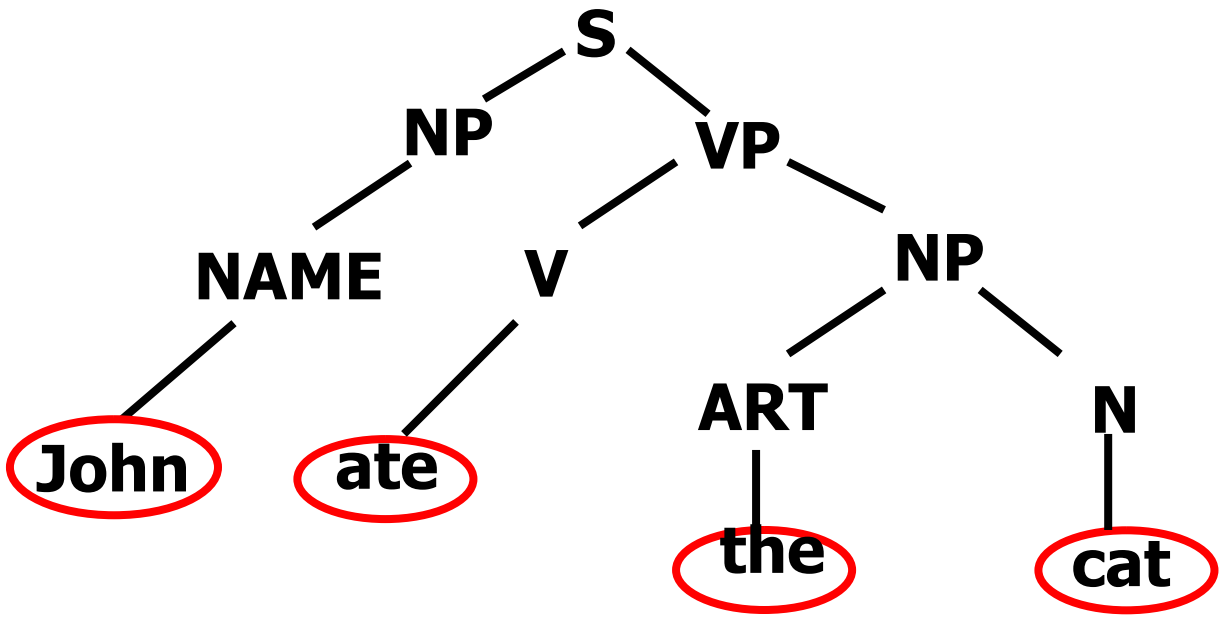
↓

包括：词根(tele-)、前缀、后缀(ing)、词尾

如like-> dislike　　boy friend-> boyfriend

1.3 自然语言处理的基本问题

句法学 **(Syntax)** 问题：研究句子结构成分之间的相互关系和组成句子序列的规则 。



1.3 自然语言处理的基本问题

语义学 **(Semantics)** 问题：研究如何从一个语句中词的意义，以及这些词在句法结构中的作用推导出该语句的意义。

这句话说了什么？

(1) 苹果不吃了。
(2) 这个人真牛。
(3) 这个人眼下没些什么，那个人嘴不太好。
(4) 火烧圆明园 / 火烧驴肉

1.3 自然语言处理的基本问题

语用学**(Pragmatics)** 问题：研究在不同上下文中的语句的应用，以及上下文对语句理解所产生的影响。

更侧重于理解其“非字面含义”。

（1）你真讨厌！

（2）A: 看看鱼怎么样了？

B: 我刚才翻了一下。

1.3 自然语言处理的基本问题

灵魂八问

配钥匙师傅：你配吗？

食堂阿姨：你要饭吗？

算命先生：你算什么东西？

快递小哥：你是什么东西？

上海垃圾分拣阿姨：你是什么垃圾？

滴滴司机：你搞清楚你自己的定位了么？

理发师傅：你自己照照镜子看看你自己，觉得还行么？

小区保安：你是谁？你从哪里来？要到哪里去？

1.4 自然语言处理面临的困难

自然语言中大量存在的歧义(ambiguity)现象

1、结构歧义

例如: (1) 喜欢乡下的孩子。

(2) 关于鲁迅的文章。

(3) 今天中午吃馒头。

(4) 今天中午吃食堂。

1.4 自然语言处理面临的困难

2、语义歧义

他说：“她这个人真有意思(**funny**)”。她说：“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(**wish**)，并让他向她意思意思(**express**)。他火了：“我根本没有那个意思(**thought**)”！她也生气了：“你们这么说是什么意思(**intention**)”？事后有人说：“真有意思(funny)”。也有人说：“真没意思(**nonsense**)”。

1.4 自然语言处理面临的困难

3、自然语言中存在未知语言现象

- **新的词汇**

  例如:“非典”、专业术语、外来语、人名等

- **新的含义**

  例如：苹果、老虎等

- **新的用法和句型**

  口语或部分网络语言中，不断出现一些“非规范的”新的语句结构，如：百度一下

1.5 NLP基本方法与技术

基本方法

- 理性主义方法 → 基于规则的方法
- 经验主义方法
- 连结主义方法 → 数据驱动的方法
  - 神经网络方法

1.5 NLP基本方法与技术

◆理性主义：通过对一些代表性语句或语言现象的研究，归纳语言使用的规律，以此分析、推断测试样本的预期结果。

- 问题求解的基本思路：基于规则的分析方法建立符号处理系统
  - 规则库开发：　N + N → NP
  - 词典标注:　#工作，N(uc)；V；
  - 推导算法设计：归约、推导、歧义消解方法…

知识库 + 推理系统 → NLP 系统

理论基础：Chomsky 的文法理论

1.5 NLP基本方法与技术

◆经验主义：利用大规真实语言数据，借助人的帮助(标注数据和筛选特征等)，统计发现语言使用的规律及其概率大小，以此预测测试样本的可能结果。统计单元是离散事件(词、短语、词性等)。

- 求解问题的思路：基于大规模真实数据建立计算模型
  - 大规模真实数据的收集、标注
  - 模型构建：模型、参数训练方法

标注语料库 + 统计模型 → NLP 系统

理论基础：统计学、信息论、机器学习

1.5 NLP基本方法与技术

◆**连结主义**：利用大规真实语言数据构建模型，统计发现语言使用的规律及概率大小，以此预测测试样本的可能结果。统计单元采用连续的实数空间表示(向量)。

- **求解问题的思路**：基于大规模真实数据建立计算模型
  - **大规模真实数据的收集**
  - **模型构建**：模型、参数训练方法

语料库 + 神经网络 + 统计模型 → NLP 系统

理论基础：统计学、深度学习

1.6 自然语言处理市场前景广阔

5.1 NLP算法工程师

北京地区平均月薪为34410。如图所示：

1.6 自然语言处理市场前景广阔

今日头条NLP算法工程师的招聘需求：

【语音识别方向】

至少在以下领域有过研究或工程经验：文本分类、知识图谱、文本挖掘、文本相似性、命名实体识别、分词、信息检索、Q&A、机器翻译；

熟悉常见NLP相关模型，如HMM、EM、LDA等；熟悉深度学习相关技术，如句向量、CNN、RNN、LSTM等模型；

熟悉 Java、C/C++、Python其中一种开发语言，有数据结构与算法的基础。

【知识图谱方向】

具备机器学习/数据挖掘理论和技术基础；

有丰富的中文NLP、QA、知识图谱、事理图谱、机器翻译、阅读理解、信号处理等项目经验，基础扎实，编码能力强；

【对话机器人方向】

熟悉NLP、机器学习、模式识别等常用算法，熟悉NLP领域当前热点和前沿技术，熟练掌握C/C++编程语言和Python，Shell等脚本语言；

有相关项目经历，包文本分类、信息抽取、知识图谱、机器学习、自动摘要等，有深度学习背景；

Thanks

谢谢！