自然语言处理（NLP）—
信息检索与机器翻译

软件学院

迟呈英　战学刚

2018年5月12日

内容提要

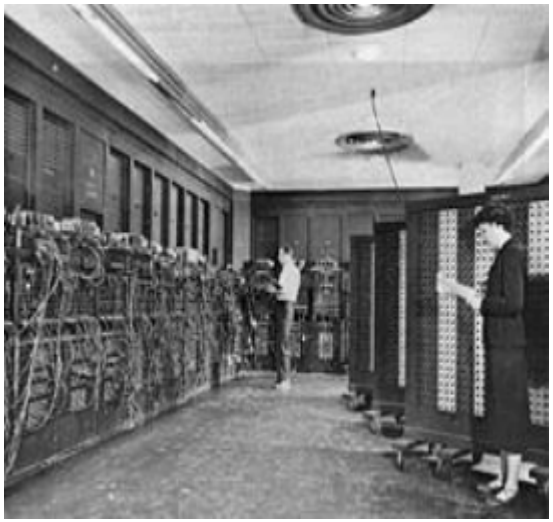
1 引言

2 NLP方法概述

3 信息检索

4 机器翻译

1 引言



Warren Weaver (July 17, 1894 –Nov. 24, 1978)

- 信息论先驱
- **1920**至**1932**年**Wisconsin**大学数学教授
- **1932**至**1955**年担任**Rockefeller Institute**自然科学部主任

1946年，世界上第一台计算机ENIAC

- **A. D. Booth** 数学物理学家，二战中参与计算机研制，在程序化计算机研究中成绩卓著；
- **1947**年**3**月至**9**月，曾在普林斯顿大学参与**John von Neumann** 研究组，后来曾在伦敦大学工作。

1 引言

达特茅斯学院**(Dartmouth College)** **(成立于1769年)**

左起：摩尔、麦卡锡、明斯基、赛弗里奇**(Oliver Selfridge)**、所罗门诺夫

人工智能夏季研讨会(大茅斯会议, **1956)**
Summer Research Project on Artificial Intelligence(Dartmouth Conference)

1 引言

●自然语言理解(naturallanguageunderstanding,NLU)是人工智能最重要的研究方向之一

●计算语言学(ComputationalLinguistics,CL)

1960S，形成相对独立的学科。1962年国际计算语言学学会(ACL)成立，1965年国际计算语言学委员会(ICCL)成立，1966年“计算语言学”首次出现在美国国家科学院ALPAC报告里

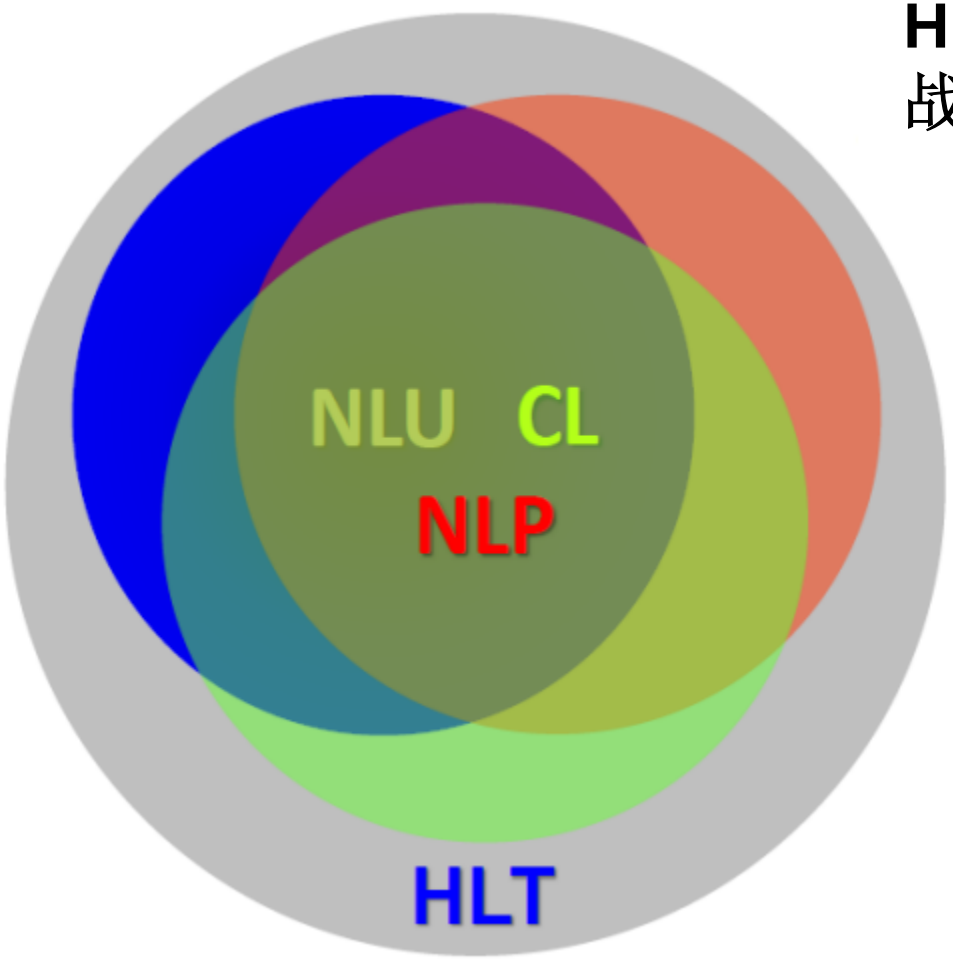
●自然语言处理(NaturalLanguageProcessing,NLP)

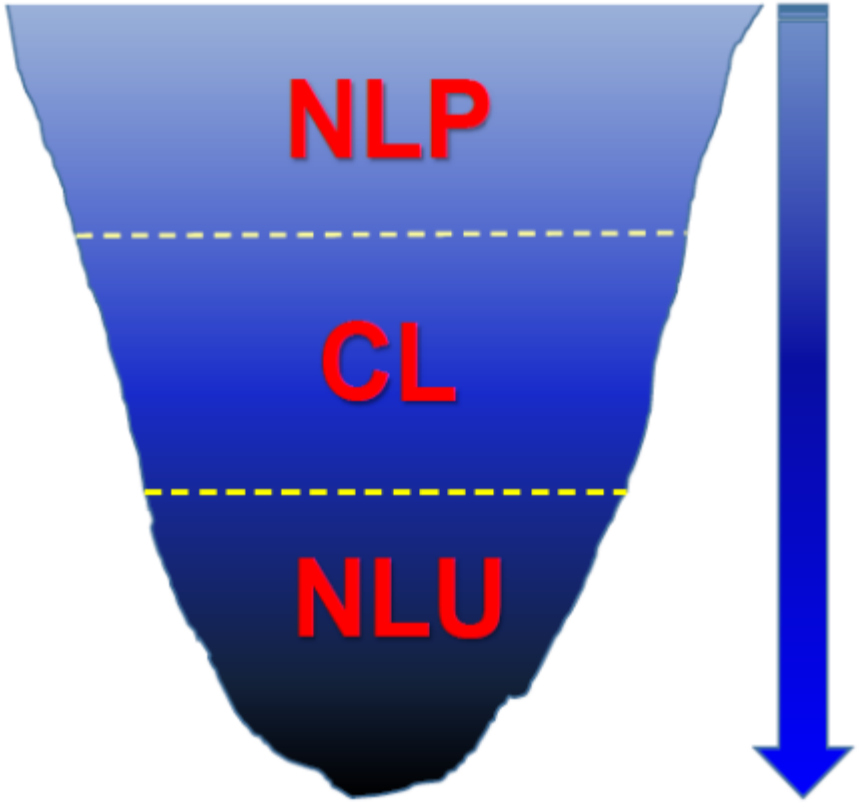
1980S，面向计算机网络和移到通信，从系统实现和语言工程的角度开展语言信息处理方法的研究。专门针对中文的语言信息技术研究成为中文信息处理

NLU、CL和NLP统称为人类语言技术(Human Language Technology, HLT)

1 引言

HLT 是当前人工智能领域最具挑战性的研究方向之一。





1 引言

网上87.8%为文本内容

移到终端：微信、短信……

非结构化文本➔语义概念关系分析、表示➔应用系统

机器翻译

情感分析

自动摘要

问答系统

观点挖掘

关系抽取

1引言

全球数万□网页

80%非汉语文字

出境游人数破亿，

前20出境游目的地

有12语言之多

64个国家和地区

44亿人口

50多种语言

1引言

Sunday August 06 -- Friday August 11, 2017

International Convention Centre, Sydney AUSTRALIA

1引言

Tuesday July 10 -- Sunday July 15, 2018

International Convention Centre，Stockholm Sweden

1引言

根据 ICML 官方的消息，2017年的最佳论文奖(Best Paper Award)被 Pang Wei Koh 和 Percy Liang 收入囊中，其中 Pang Wei Koh 目前是斯坦福大学的在读博士生，而 Percy Liang 则是斯坦福大学的助理教授，都是华人。

1 引言

问题与挑战

- 大量的未知现象
  如：高山, 埃博拉，奥特
- 无处不在的歧义词汇
  如：苹果,粉丝,Bank
- 复杂或歧义结构比比皆是
  喜欢乡下的孩子。Time flies like an arrow.
- 普遍存在的缩略和隐喻表达
  要把权力装进制度的笼子；老虎苍蝇一起打。
  破四旧，除四害；消灭一切牛鬼蛇神。

1引言

问题与挑战

●跨语言语义概念不对等

如：馒头:steamed bread

We do chicken right.

我们对鸡做的权利。(Google Translate，2018.5.5)

我们是烹鸡专家。(百度翻译, 2018.5.4.)

NLP要解决的问题是从大量不确定性中寻找确定性结论，很多背景知识和常识性知识是隐含的，是在语义和概念层面上进行的表示、处理和变换。

2 NLP方法概述

2.1基本方法

●**理性主义方法**：1957～1980S

* 词法分析，句法方法，语义分析
* 词典、规则 - **基于规则的方法**

●**经验主义方法**：～1950S，1980S～

* 训练样本
* 统计模型 - **基于统计的方法**

2 NLP方法概述

- 以机器翻译为例

给定英语句子：

There is a book on the desk.

将其翻译成汉语。

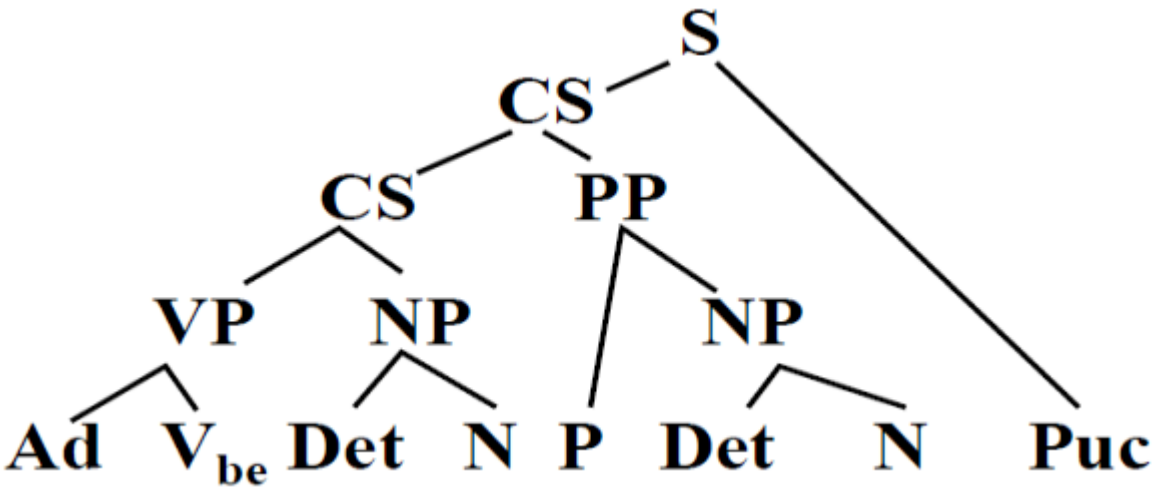
2 NLP方法概述

➢基于规则的方法

◆对英语句子进行词法分析

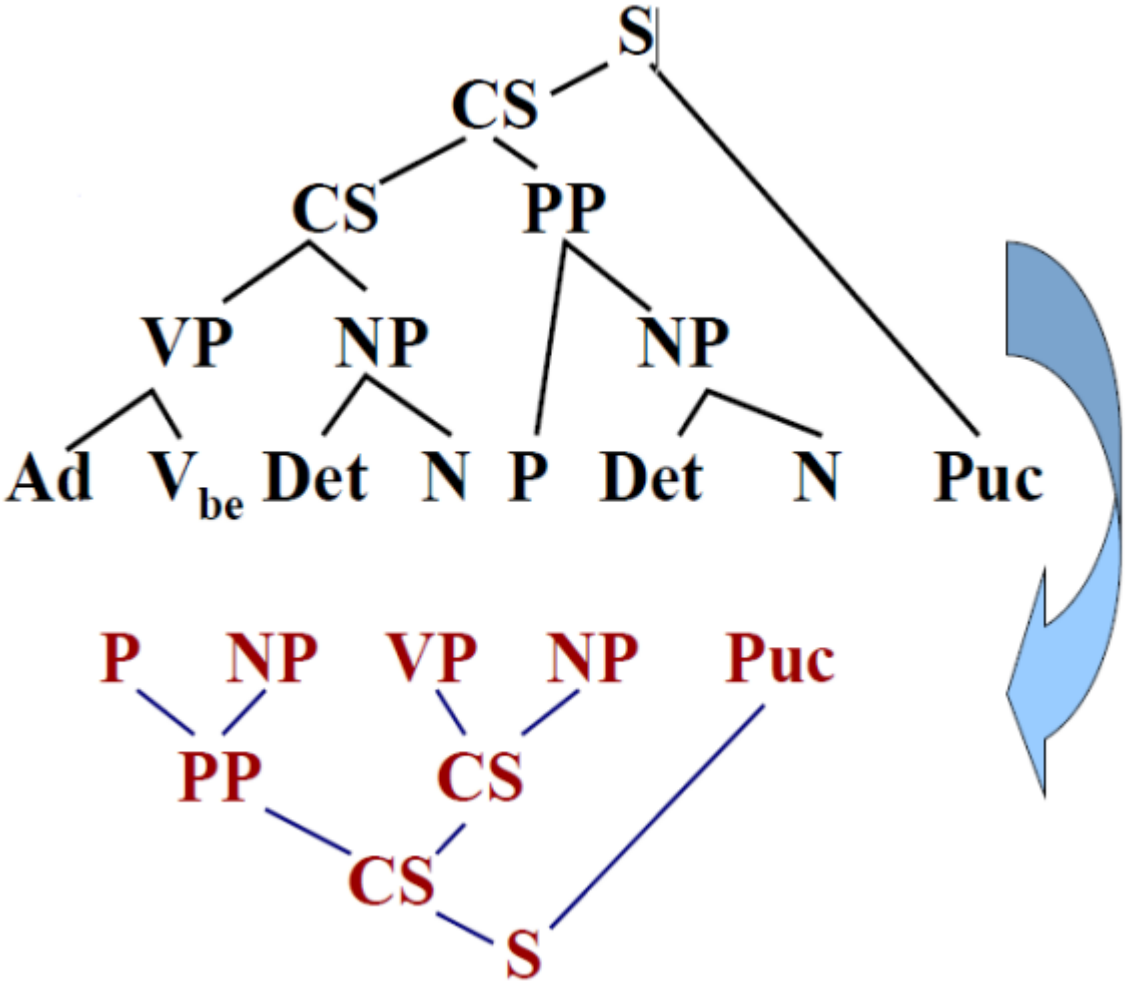
There/A_d is/V_{be} a/Det book/N on/P the/Det desk/N./Puc

◆对英语句子进行句法结构分析



2 NLP方法概述

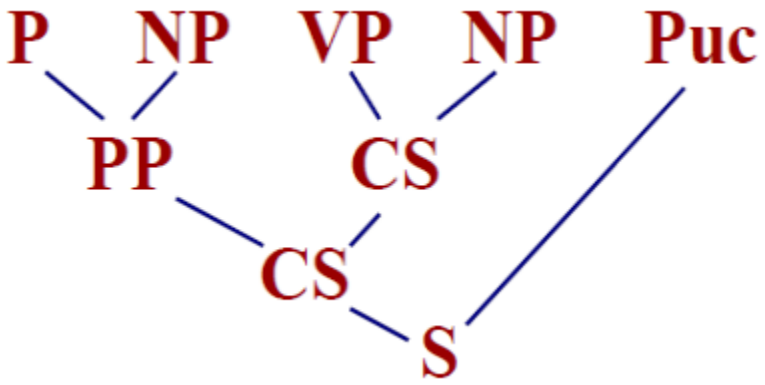
◆利用转换规则将英语句子结构转换成汉语句子结构



2 NLP方法概述

◆根据转换后的句子结构，利用词典和生成规则生成翻译的结果句子

#a，Det, 一
#book，N，书；V，预订
#desk，N，桌子
#on，P，在X 上
#There be，V, 有



输出译文：

在桌子上有一本书。

基于规则的**NLP方法的基本步骤**：
词法分析(汉语分词) → 句法分析 → 语义分析(词义消歧等) → 语言生成

2 NLP方法概述

➢基于统计的方法

给定源语言句子：$E = e_1^m \equiv e_1 e_2 \cdots e_m$

将其翻译成目标语言句子：$C = c_1^l \equiv c_1 c_2 \cdots c_l$

根据贝叶斯公式：$P(C|E) = \dfrac{P(C) \times P(E|C)}{P(E)}$

$$\hat{C} = \arg\max_{C} P(C) \times P(E|C)$$

语言模型
(Language model, LM)

翻译模型
(Translation model, TM)

2 NLP方法概述

构建解码器(decoder)，快速搜索最优翻译候选：

◆三个关键问题：

- 估计语言模型概率 $p(C)$；
- 估计翻译模型概率 $p(E|C)$；
- 快速有效地搜索候选译文C，使$p(C)\times p(E|C)$最大。

◆主要任务：

- 收集大规模双语句子对、目标语言句子
- 参数训练与模型优化

2 NLP方法概述

基本原理：一个句子是否合理，就看他的**可能性大小**如何。概括来说：假定S表示某一个有意义的句子，由一连串特定顺序排列的词$w_1,w_2,...,w_n$组成，这里的n表示句子的长度。则概率P(S)表示上面句子的合理性。$P(S)=P(w_1,w_2,...,w_n)$

利用条件概率公式：

$$P(w_1,w_2,...,w_n)=P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)...P(wn|w_1,w_2,...,w_{n-1})$$

上式中，$P(w_1)$表示第一个词出现的概率，$P(w_2|w_1)$是在已知第一个词的前提下，第二个词出现的概率，以此类推。简单的看一下上面的公式，可以发现除了$P(w_1)$以及后面的$P(w_2|w_1)$比较好算以外，其他的项计算难度都比较大。

俄国科学家马尔科夫给出了一个假设——假设任意一个词w_t出现的概率只同它前面的词w_{t-1}有关。

于是上面的公式就可以简化为：

$$P(w_1,w_2,...,w_n)=P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$$

- **贝叶斯公式：** $P(w_i|w_{i-1})=P(w_{i-1},w_i)/P(w_{i-1})$

  其中$P(w_{i-1},w_i)$可以样本的相对频率(样本数量足够)来统计。具体公式如下：

- $P(w_{i-1},w_i)=N(w_{i-1},w_i)/N(w_{i-1})$

  其中$N(w_{i-1},w_i)$代表在样本中w_{i-1},w_i和前后相邻出现了多少次。$N(w_{i-1})$表示在样本中w_{i-1}出现了多少次。

2 NLP方法概述

人类共有二十三对染色体。humans have a total of 23 pairs of chromosomes .

澳洲重新开放驻马尼拉大使馆 australia reopens embassy in manila

中国大陆手机用户成长将减缓 growth of phone users in mainland china to slow

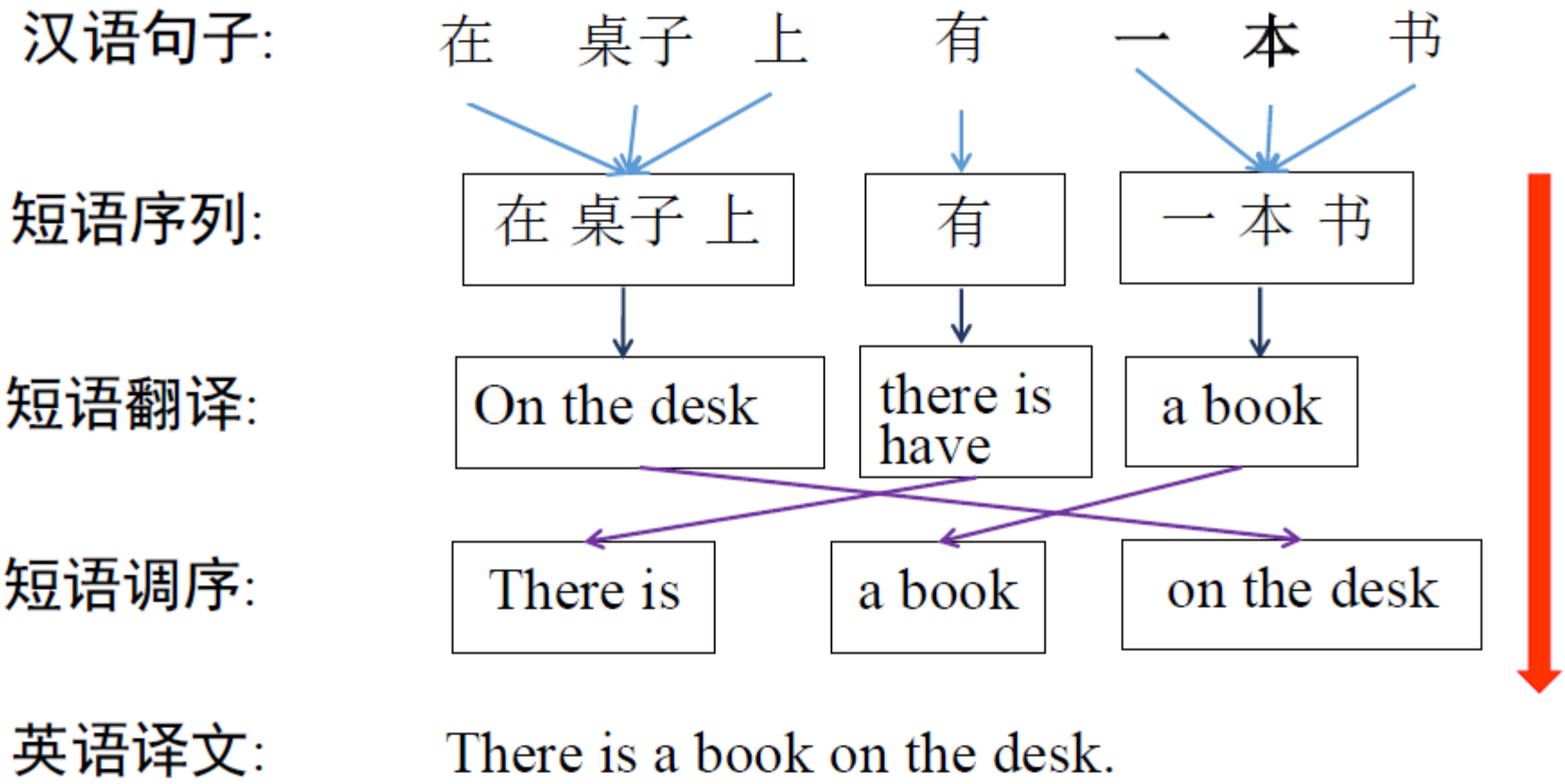
外交人员搭乘第五架飞机返国 diplomatic staff will take the fifth plane home .

驻南韩美军三千人奉命冻结调防 us freezes transfer of 3,000 troops in south

korea姚明感慨NBA 的偶像来得太快 yao ming feels nba stardom comes too fast

双语句对

2 NLP方法概述


汉语句子:
在 桌子 上 有 一 本 书
短语序列:
在 桌子 上
有
一 本 书
短语翻译:
On the desk
there is
have
a book
短语调序:
There is
a book
on the desk
英语译文:
There is a book on the desk.

2 NLP方法概述

2.2常用的统计模型和开源工具

- 感知机(perceptron)：二类分类
- *k-近邻法(k-nearestneighbor,k-NN)：多类分类问题*
- 朴素贝叶斯法(naïveBayes)：多类分类问题
- 决策树(decisiontree)：多类分类问题
- 最大熵(maximumentropy)：多类分类问题
- 支持向量机(supportvectormachine,SVM)：二类分类
- 条件随机场(conditionalrandomfield,CRF)：序列标注
- 隐马尔可夫模型(hiddenMarkovmodel,HMM)：

2 NLP方法概述

开源工具：

●条件随机场：

◆**CRF++（C++版）：**

http://crfpp.googlecode.com/svn/trunk/doc/index.html

◆**CRFSuite（C语言版）：**

http://www.chokkan.org/software/crfsuite/

◆**MALLET (Java版，通用的NLP工具包，包括分类、序列标注等机器学习算法)：**http://mallet.cs.umass.edu/

◆**NLTK (Python版，通用的NLP工具包，很多工具是从MALLET中包装转成的Python接口)：**http://nltk.org/

2 NLP方法概述

- **贝叶斯分类器**：http://www.openpr.org.cn
- **支持向量机**(LibSVM)：
  http://www.csie.ntu.edu.tw/~cjlin/libsvm
- **隐马尔可夫模型**：http://htk.eng.cam.ac.uk/
- **最大熵**：
  - **OpenNLP**：http://incubator.apache.org/opennlp/
  - **Malouf**：http://tadm.sourceforge.net/
  - **Tsujii**：http://www-sujii.is.s.utokyo.ac.jp/~tsuruoka/maxent/
  - 张乐：http://homepages.inf.ed.ac.uk/lzhang10/maxent.html
  - **林德康**：http://webdocs.cs.ualberta.ca/~lindek/downloads.htm

3 信息检索

➢**信息检索 (information retrieval , IR)**是指将信息按一定的方式组织和存储起来，并根据用户的需要找出有关信息的过程。**1950**年**,** 穆尔**(Moore C)**根据图书馆的参考咨询和文摘索引提出了信息检索。

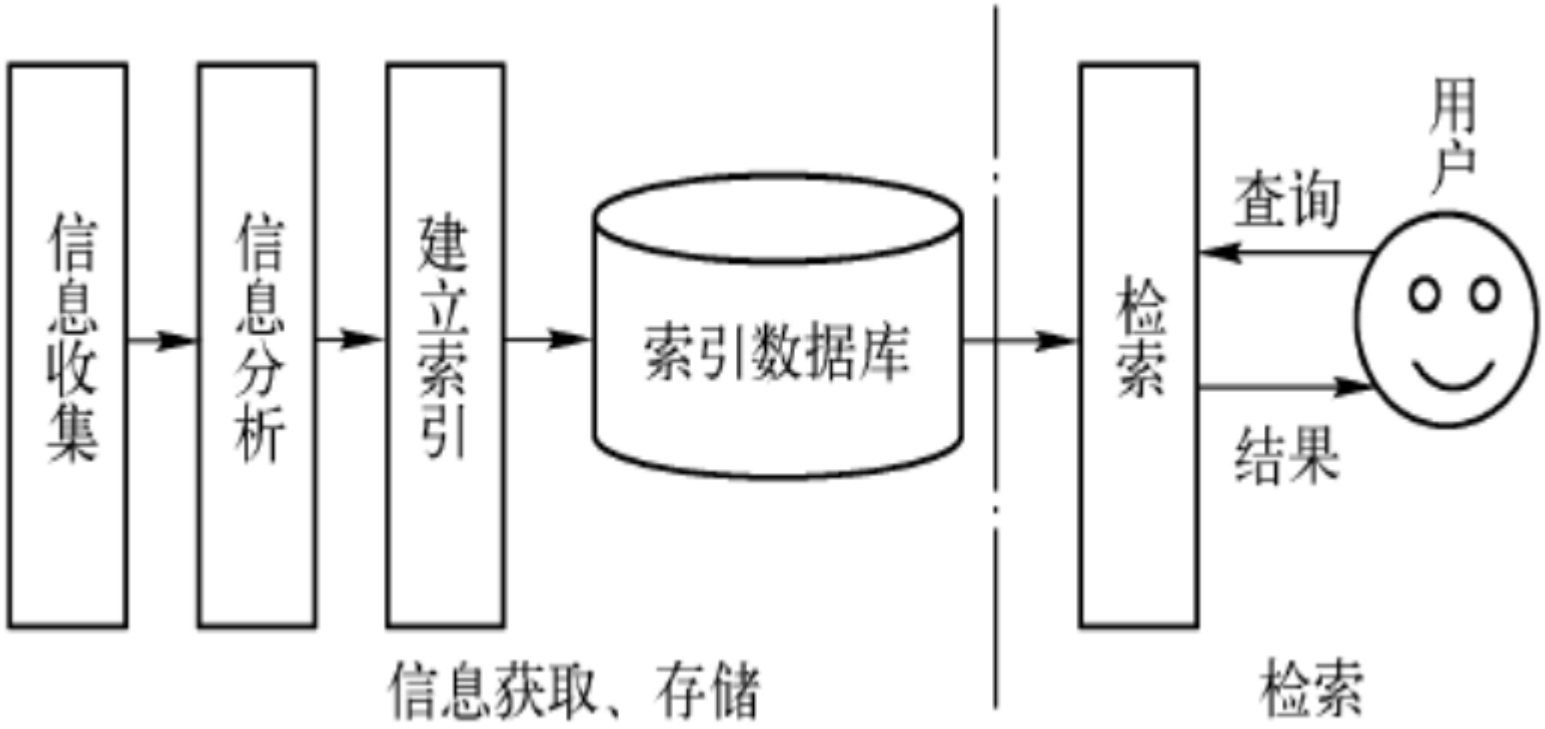
➢**信息检索包括信息存储和检索。**

◆在检索之前必须将信息收集起来，按科学方法进行整理，并按一定准则存储起来，形成书本式检索工具或者计算机可读数据库。

◆在检索时，用户根据自身需求提交查询给信息检索系统，系统利用存储信息所依据的准则，在文档集中找出与查询条件相关的文档子集，并按照它们与查询条件的相关性进行排序。

◆最后为用户返回一个有序的文档子集。

3 信息检索



信息检索模型

3 信息检索

➢**按信息检索的内容划分**

文献检索
数据检索
事实检索
概念检索

➢**按信息检索的组织方式划分**

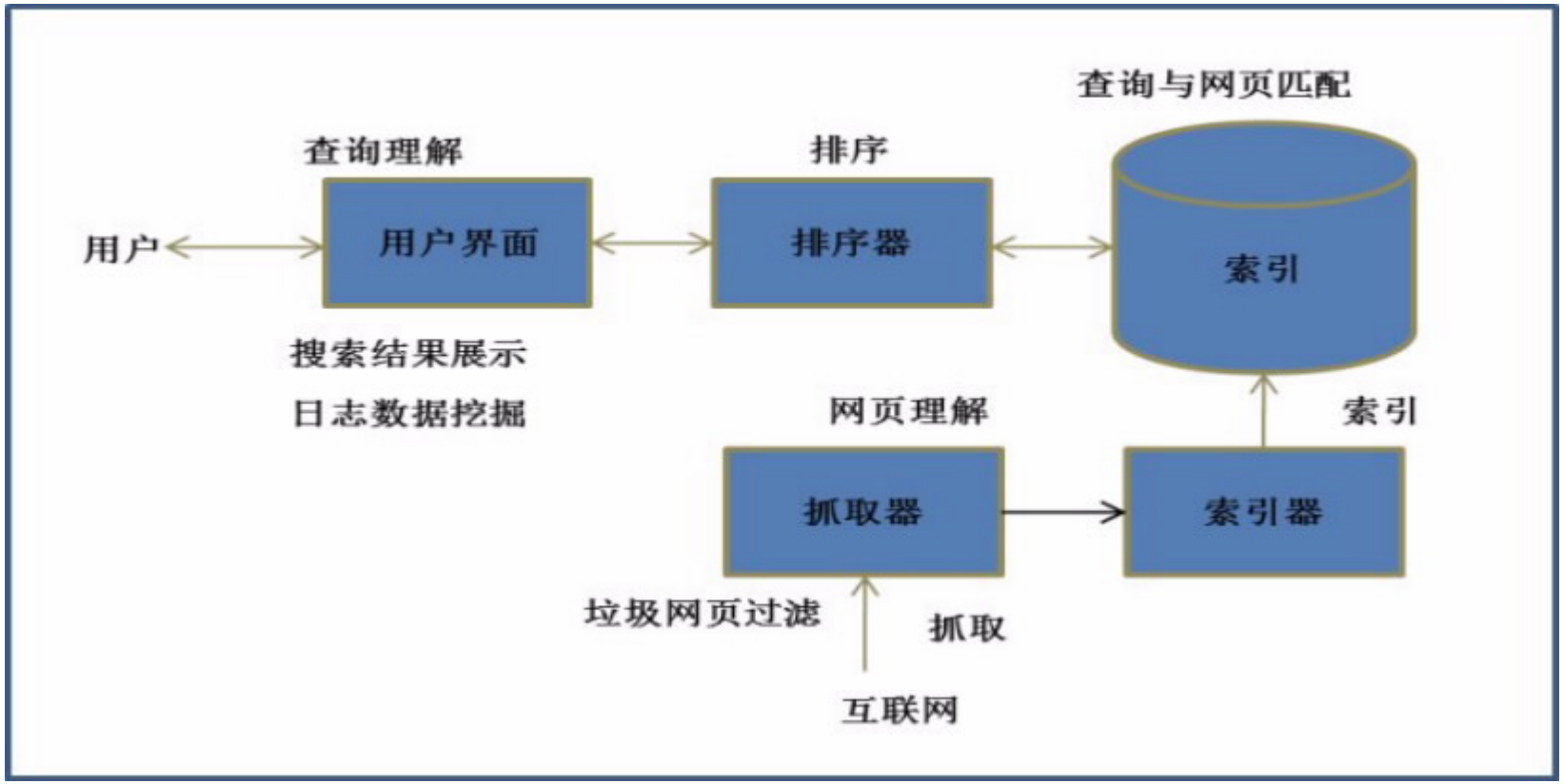
- 全文本检索
- 多媒体检索
- 超文本检索

超文本检索是对每个节点中储存的信息以及信息链构成的网络信息进行的检索。与传统文本的线性顺序不同，超文本检索强调中心节点之间的语义联结结构，靠系统提供的工具进行图示穿行和节点展示，提供浏览式查询，可进行跨库检索。

3 信息检索

据不完全统计，约有60%的互联网用户每天至少使用一次搜索引擎，约有90%的互联网用户每周至少使用一次搜索引擎。搜索引擎已经成为人们访问互联网的必经通道。对一般用户来说，除了搜索，没有其他手段可以帮助更方便地获得网上的信息。如果图灵在世的话，他会惊喜地发现互联网搜索引擎已经能在自己当年设计的人工智能测试上取得相当好的成绩，因为在主要的搜索引擎上提出各种各样的问题，比如“理想国的作者？”或者“从知春路到清华东门怎么坐公交车？”，都能找到正确的答案。毫无疑问，互联网搜索引擎已成为当今最为实用、最具代表性的智能系统。

3 信息检索



互联网搜索引擎架构图

3 信息检索

➢ 互联网搜索经历了三代的发展历程：

- **第一代搜索：**

  将互联网网页看作文本，采用传统信息检索的方法。

- **第二代搜索：**

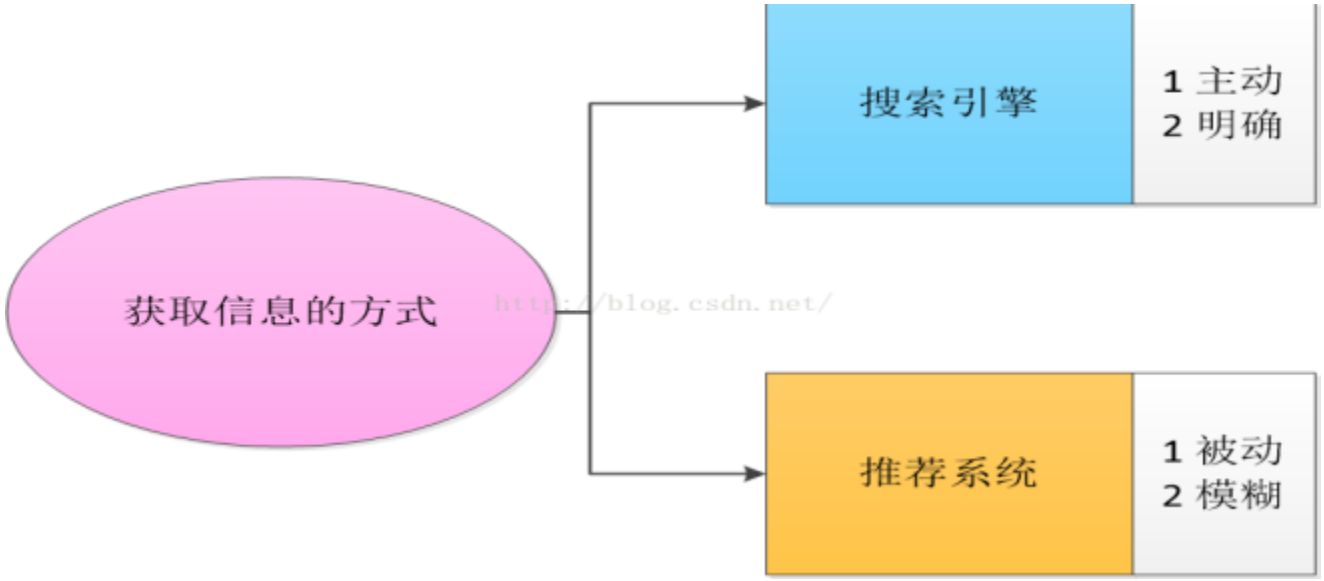
  利用互联网的超文本结构，有效地计算网页的相关度与重要度，代表的算法有PageRank。

- **第三代搜索：**

  有效利用日志数据与统计学习，使网页相关度与重要度计算的精度有了进一步的提升，代表的技术包括排序学习、网页重要度学习、匹配学习、话题模型学习、查询语句转化学习。

3 信息检索

但是随着计算机技术的发展，做搜索的一般也做推荐，从信息获取的角度来看，目前搜索和推荐是用户获取信息的两种主要手段。无论在互联网上，还是在线下的场景里，搜索和推荐这两种方式都大量并存。



3 信息检索

➢搜索引擎

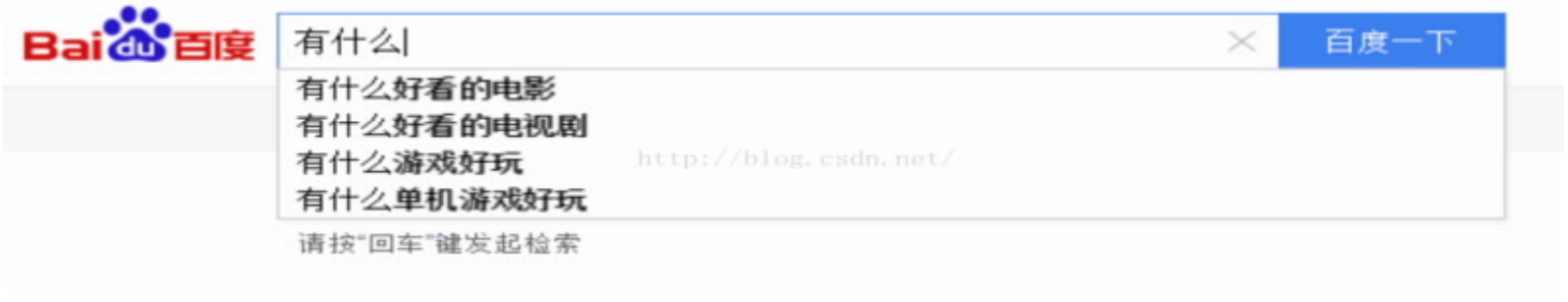
获取信息是人类认知世界、生存发展的刚需，搜索就是最明确的一种方式，其体现的动作就是“出去找”，找食物、找地点等，到了互联网时代，搜索引擎（Search Engine）就是满足找信息这个需求的最好工具，你输入想要找的内容，搜索引擎快速的给你最好的结果，这样的刚需催生了Google、百度这样的互联网巨头。

➢推荐系统

但是获取信息的方式除了搜索外，还有另一类，称为推荐系统（Recommendation System,简称Recsys），推荐也是伴随人类发展而生的一种基本技能，你一定遇到这样的场景，初来乍到一个地方，会找当地的朋友打听“嗨，请推荐下附近有啥好吃好玩的地方吧！”——知识、信息等通过推荐来传播，这也是一种获取信息的方式。

3 信息检索

搜索和推荐的区别如下图所示，**搜索**是一个非常**主动**的行为，并且**用户的需求十分明确**，在搜索引擎提供的结果里，用户也能通过浏览和点击来明确的判断是否满足了用户需求。然而，**推荐系统**接受信息是**被动**的，**需求也都是模糊而不明确的**。以“逛”商场为例，在用户进入商场的时候，如果**需求不明确，这个时候需要推荐系统**，来告诉用户有哪些优质的商品、哪些合适的内容等，但如果**用户已经非常明确**当下需要购买哪个品牌、什么型号的商品时，直接去找对应的店铺就行，这时就是搜索了。



3 信息检索

搜索和推荐虽然有很多差异，但两者都是大数据技术的应用分支，存在着大量的交叠。近年来，搜索引擎逐步融合了推荐系统的结果，例如右侧的“猜你关注”、下一页底部的“相关搜索”等，都使用了推荐系统的产品思路和运算方法（如下图红圈区域）。

3 信息检索


www.xiachufang.com>全部分类 - 快照 - 下厨房
海参的家常哪些人适合吃海参?怎么吃海参好?
广告
海参的家常哪些人适合吃海参?怎么吃海参好?在线参农为您解答!海参的家常海参,品质好,价格优,吃过的都说效果好,欢迎咨询!
bmf.etcom162.cn - 2018-05-06
长兴岛老黄海参 没有中间商 高品质纯淡干野生海参
纯淡干野生，自己深海捕捞加工，刺多肉厚，追求高品质纯天然野生海参的营养功效！品质佳口碑好 好多客户都是老客户介绍的
360.qiushei.cn/428/hs2600/ - 2018-05-06
海参怎么样好 一天吃多少为宜?注意什么 你知道吗?
海参怎么样好 海参一种营养价值高的海珍品,不含胆固醇,蛋白质含里高,味甘,咸,性温,自古以来,陆有人参,水有它海参怎么样好,两参齐名
hs.553027.top - 2018-05-06
相关搜索 | 反馈
海参的家常做法
葱烧海参
干海参怎么吃法
海参怎么吃
海参吃法大全
海参小米粥
海参的价格
海参
海参怎么吃最有营养
海参的功效与作用
海参怎么泡发
冷冻海参的吃法
干海参
即食海参的吃法
干海参的吃法
1 2 3 4 5 6 7 8 9 10 下一页
找到相关结果约2,680,000个

3 信息检索

推荐系统也大量运用了搜索引擎的技术，搜索引擎解决运算性能的一个重要的数据结构是倒排索引技术（Inverted Index），而在推荐系统中，一类重要算法是基于内容的推荐（Content-based Recommendation），这其中大量运用了倒排索引、查询、结果归并等方法。另外点击反馈（Click Feedback）算法等也都在两者中大量运用以提升效果。

3 信息检索

目前指导的研究生研究基于用户画像的推荐，用户画像在大数据分析中是一种很有用的系统，它可以各种不同的系统中，起到很关键的作用。比如搜索引擎、推荐系统等等，可以帮助应用实现千人千面、个性化、精准等的效果。

- **搜索引擎**

在搜索的时候考虑用户的画像标签，返回用户感兴趣的内容。比如同一个关键字“诸葛亮”，王者荣耀的爱好者搜索的时候应该返回“诸葛亮”相关的内容，比如如何加铭文、如何五杀等等；而历史爱好者搜索的时候，应该返回三国相关的内容。

- **推荐系统**

推荐系统可以根据用户的喜好和特征，也就是用户的画像，推荐相关的内容。比如，给一个用户定位的画像是美妆达人，那么就应该给她多推送一些面膜护肤之类的东西，而不是推一堆零食。

4 机器翻译

➢**机器翻译 (machine translation, MT)**

利用计算机把一种语言(源语言, source language)翻译成另一种语言(目标语言, target language) 的一门学科和技术。

➢**机器翻译的产生发展**

◆**1947 ~ 1964：开创时期**

◆**1970 ~ 1976：复苏阶段**

◆**1976 ~ 现在：繁荣时期**

4 机器翻译

1．开创期（1947-1964)

1954 年，美国乔治敦大学在 IBM 公司协同下，用IBM-701计算机首次完成了英俄机器翻译试验，拉开了机器翻译研究的序幕。

中国早在1956年,国家就把这项研究列入了全国科学工作发展规划，课题名称是“机器翻译、自然语言翻译规则的建设和自然语言的数学理论”。

从20世纪50年代开始到20世纪60年代前半期，美国和前苏联两个超级大国出于军事、政治、经济目的，均对机器翻译项目提供了大量的资金支持，而欧洲国家由于地缘政治和经济的需要也对机器翻译研究给予了相当大的重视，机器翻译一时出现热潮。

4 机器翻译

2．受挫期（1964-1975）

1964年，为了对机器翻译的研究进展作出评价，美国科学院成立了语言自动处理咨询委员会(Automatic Language Processing Advisory Committee，简称ALPAC委员会)，开始了为期两年的综合调查分析和测试。

1966年11月，该委员会公布了一个题为《语言与机器》的报告（简称ALPAC报告），该报告全面否定了机器翻译的可行性，并建议停止对机器翻译项目的资金支持。这一报告的发表给了正在蓬勃发展的机器翻译当头一棒，机器翻译研究陷入了近乎停滞的僵局。无独有偶，在此期间，中国爆发了“十年文革”，基本上这些研究也停滞了。机器翻译步入萧条期。

4 机器翻译

3．恢复期（1975-1989）

进入 70 年代后，随着科学技术的发展和各国科技情报交流的日趋频繁，国与国之间的语言障碍显得更为严重，传统的人工作业方式已经远远不能满足需求，迫切地需要计算机来从事翻译工作。同时，计算机科学、语言学研究的发展，特别是计算机硬件技术的大幅度提高以及人工智能在自然语言处理上的应用，从技术层面推动了机器翻译研究的复苏，机器翻译项目又开始发展起来，各种实用的以及实验的系统被先后推出，例如 Weinder 系统、EURPOTRA 多国语翻译系统、TAUM-METEO系统等。

4 机器翻译

4．新时期（1990～现在）

随着 Internet 的普遍应用，世界经济一体化进程的加速以及国际社会交流的日渐频繁，传统的人工作业的方式已经远远不能满足迅猛增长的翻译需求，人们对于机器翻译的需求空前增长，机器翻译迎来了一个新的发展机遇。国际性的关于机器翻译研究的会议频繁召开，中国也取得了前所未有的成就，相继推出了一系列机器翻译软件，例如“译星”、“雅信”、“通译”、“华建”等。在市场需求的推动下，商用机器翻译系统迈入了实用化阶段，走进了市场，来到了用户面前。

4 机器翻译

基本翻译方法

- 直接转换法
- 基于规则的翻译方法
- 基于中间语言的翻译方法
- 基于语料库的翻译方法
  - 基于实例的翻译方法
  - 基于统计的翻译方法

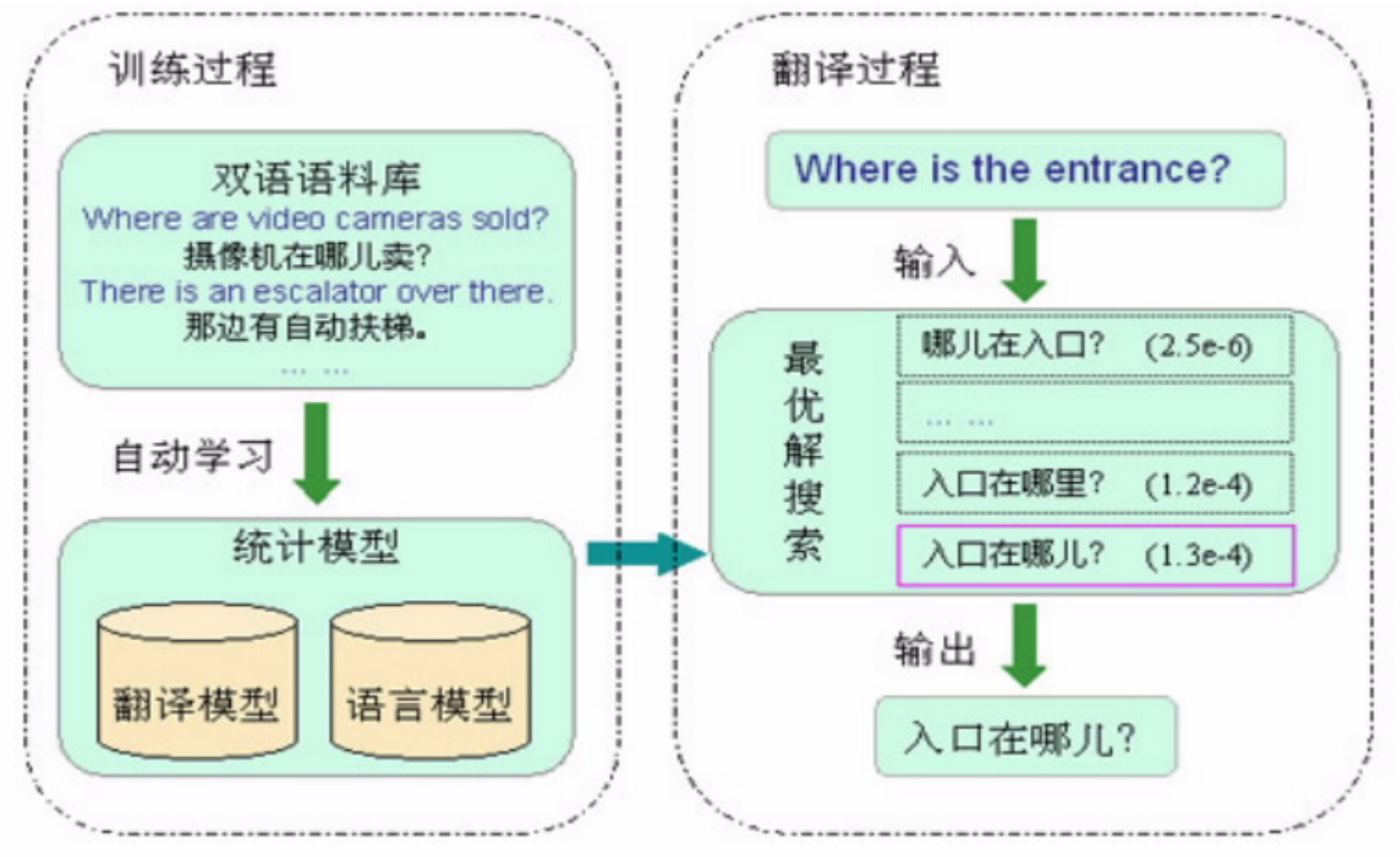
4 机器翻译

基于统计的机器翻译方法

- 统计机器翻译的**基本思想**是通过对大量的平行语料进行统计分析，构建**统计翻译模型**，进而使用此模型进行翻译。

- 通俗地说，源语到目的语的翻译是一个**概率统计问题**，任何一个目的语句子都有可能是任何一个源语句的译文，只是概率不同，**机器翻译的任务就是找到概率最大的句子**。

4 机器翻译



统计机器翻译模型

4 机器翻译

优势

◆更好地利用资源

（1）存在着大量可被机器读取的格式的自然语言。

（2）通常，统计机器翻译系统不是针对于任何具体的语言配对。

（3）基于规则的翻译系统需要对语言规则的手动开发，这样不仅成本很高，而且通常对其它语言不适用。

◆更多的自然语言翻译资料

（1）文档的翻译在基于可能性的p(e | f)，其中的本国语言（例如英语）字符“e” 就是对外国语言（例如法文）中字符“f”的翻译。这些可能性都是利用参数估算的技术实现。

（2）将贝叶斯法则应用于p(e | f)，会得这个外语字符译成母语字符的可能性，从算术上来说，发现最佳译文也就是选取出现概率最高的那个。

4 机器翻译

统计机器翻译模型分类

➢基于词汇的的翻译：

在基于词汇的翻译中，待译的元素是所有的词汇。实际上，由于复合的词汇、词法和习惯用语，待译语句中的词汇数量是不同的。所译词汇顺序的长度比被称作生产率，它告诉我们每个母语词汇会产生多少外语词汇。简单地基于词汇的翻译无法翻译生产率不是 1 的语言对儿。要使基于词汇的翻译系统能够有效处理，例如，高生产率的情况，系统可以将一个词与多个词对应，但反之则并非如此。

基于词汇的翻译系统的一个实例是包含IBM模型的免费软件：

GIZA++package(GPLed)。

4 机器翻译

➢**基于短语的翻译模型**

短语是任何**连续的词串**。如果利用该模型将“The weather is nice today”译为中文，具体过程就是：

首先是翻译模型。实质就是学习“**对应关系**”，而每种对应关系都具备一定的概率。假设： the weather is=天气，概率是0.001 nice=很好，概率是0.2 today=今天，概率是0.5，如果到此为止，译文就是“天气很好今天”，语序显然不对。这时就要用到**语言模型**和**调序模型**。

语言模型在单语语料库中学到，“今天”有一定概率放在“天气”前面，“天气”又有一定概率放在“很好”前面。于是，结合从翻译模型、语言模型与调序模型得到的概率值，利用对数线性模型计算出一个最终得分，使机器翻译解码器认为“今天天气很好”为最优答案，从而完成翻译解码过程。

统计机器翻译的三个关键问题：（1）估计翻译模型、语言模型和调序模型的概率；（2）估计各模型的权重；（3）快速有效地搜索目标句子T。由此可见驱使这三种模型输出结果的是统计数据，而非语法规则。

4 机器翻译

基于句法的翻译对统计机器翻译的挑战，我们跟东北大学合作申请的国家自然科学基金面上项目--基于深度句法的统计机器翻译方法研究

统计机器翻译不得不处理的问题包括：

复合词

习惯用语

词法

迥异的词序

不同语言的词序也不相同。我们可以通过一个句子中的主语（S）、谓语（V）和宾语（O）的顺序来给语言分类，例如，SVO或者VSO语言。词序上还有其它的不同之处，例如，名词修饰语的位置。

4 机器翻译

基于深度句法的统计机器翻译方法研究-正在研究的课题

➢多层次深度句法表示

◆层次1：基于传统句法的转换。由于传统句法树是从单语句法分析的角度定义的，因此可以通过自动转换将其变为更加适合机器翻译的形式。

例如，图1最上层的 IP(NP VP NP)结构，被转换为IP(NP VP*(VP NP))，其中VP*是新引入的节点，这么做可以使这个句子更符合英文翻译的结构，因为英文中通常将动词和其对应的宾语从句作为一个结构翻译。此外，这种较“深”的树结构也有助于解决句法结构的稀疏问题。

4 机器翻译

◆层次2：词汇功能描述。图 1右侧中所描述的句法结构与左侧的另一处不同在于词汇描述更加丰富，比如，它描述了一个词能否接一个从句结构、能否被翻译为空串。而以上这些信息在传统的句法表示中并没有体现，但是它们对机器翻译的从句结构调序和空翻译等问题是非常重要的。

◆层次3：句法功能描述。类似于词汇功能的描述，我们也可以用更加丰富的句法功能描述对句法树中的每个节点（或者子树片段）进行定义。如图1所示，我们可以利用单词以及句法片段之间相互依赖关系以及中心动词搭配关系来进一步描述句法结构，这样可以使机器翻译系统能够使用这种关系（如主谓）进行结构调序或主谓一致性的生成。

4 机器翻译

◆层次4：翻译外部知识。人翻译一个句子时通常以一定的先验知识做指导，比如，我们翻译一个复杂句时，会先翻译句子的核心结果，之后翻译从属成分。我们把这种知识称为翻译外部知识。在图1 中，用红色（加key 标志）的部分描述了上面这种核心结构。可以看出，这种结构会决定整个英文译文的基本结构，而其它成分的翻译都附着在核心结构翻译的周围。

◆层次5：隐含向量表示。除了语言学驱动的描述形式，还可以进一步使用统计学上的描述。比如，我们可以把一个词或者句法结构抽象为一个向量，用向量中的多个维度描述一个词或者句法结构的某种隐含属性，而这种属性通过人工是不易捕捉的。本项目将重点研究基于神经元网络的句法隐含向量学习。这种自动学习到的向量可以进一步丰富一个词或者句法结构在机器翻译中的描述能力，比如定义更多维的特征提高翻译模型的判别能力。

4 机器翻译

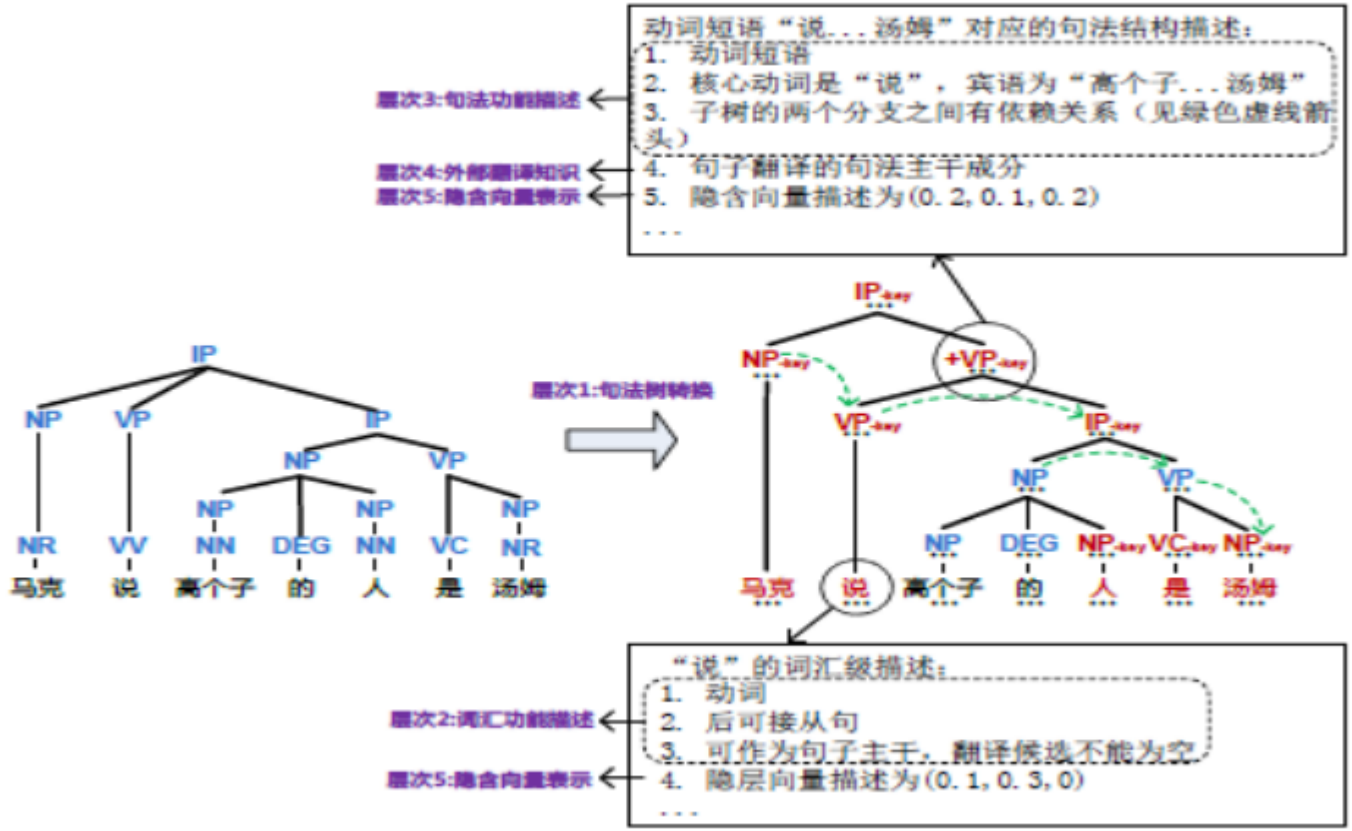


图1 多层次深度句法表示实例

4 机器翻译

➢基于深度句法的规则推导翻译模型

利用基于规则推导的方式描述翻译过程，也就是把基于句法的翻译分解为一系列的翻译规则，之后通过规则的组合拼装出符合源语和/或目标语句法的翻译。

主要方法是根据深度句法信息的层次信息，研究源语深度句法树到目标语串翻译的多层建模，内容涉及：词汇及句法功能描述翻译子模型、外部翻译知识匹配子模型、双语隐含向量翻译子模型。同时结合传统句法翻译的基础模型组合上述子模型。此外，由于以上模型及特征之间具有复杂的联系，可能具有非线性关系。因此使用神经元网络中的非线性变化以及Boosting 学习等技术实施有效的数学建模方式。

小结

- 自然语言处理远不是人们原来想象的那么简单，而是十分困难的。从现有的理论和技术现状看，通用的、高质量的自然语言处理系统，只能是长期的努力目标。
- NLP的未来既有乐观的一面，也有悲观的 一面：
  - 乐观：随着机器性能的不断提高，及各种统计方法的实践应用和完善，NLP系统的性能会不断提高。
  - 悲观：自然语言超出了目前计算机模型（Turing Machine）所能处理的范围，相关问题的解只能是近似解。

Thank You