第12讲 NLP专题之(三)

机器翻译和自动问答
machine translation
and question answering

内容

1. 词典和语料库
2. 机器翻译
3. 自动问答

一　词典及语料库

词典相关的研究

- 词典与词典编撰的研究
  - 词典学lexicology
  - 计算词典学computational lexicology
  - 词典编撰学lexicography
  - 计算词典编撰学computational lexicography

机读词典与人读词典

- 人读词典
  - – 格式不规范
  - – 数据完整性和一致性不好
  - – 非结构化
- 机读词典
  - – 格式规范
  - – 数据完整性和一致性较好
  - – 结构化

人读词典例

- 金山词霸
  - (1) 故事，小说；传闻； 轶事
  - (2) (书籍、电影、戏剧等的）情节
  - (3)（报刊、杂志文章的）素材，题材

机读词典的分类

- 按信息类型分类
  - – 语法词典
  - – 语义词典（包括同义词典）
  - – 双语词典
  - – ……
- 按领域分类
  - – 通用词典
  - – 专业词典（术语词典）
  - – 专名词典
  - – ……

例1：汉语语法信息词典

- 开发单位：北京大学计算语言学研究所
- 规模：7万多词条

词语	全拼音	同形	前时	后时	在时	到时	等到	时态	兼类	备注
晚秋	wan3qiu1			否						
晚上	wan3shang5									
往常	wang3chang2		否	否		否	否	过		
往后	wang3hou4				否	否	否	未		
往年	wang3nian2		否			否	否	过		"三月天气还很冷
未来	wei4lai2		否	否			否	未		

例2：新华社词语数据库

全库分为中文和外文两个大类，主要包括中文新闻库、经济信息库、证券库、人物库、组织机构库、专题资料库等中文数据库，还包括Xinhua News Bulletin 等英文数据库，共有28 个库100多个子库，数据量达80多亿汉字， 并以日均150万汉字的速度增长。

例3：知网

- 作者：董振东董强
- 网站：http://www.keenage.com
- 概念描述举例
- NO.=017144
  - **W_C=打**
  - **G_C=V**
  - **E_C=~网球，~牌，~秋千，~太极，球~得很棒**
  - **W_E=play**
  - **DEF=exercise|锻练,sport|体育**
  - **其中DEF是核心，采用特定的“知识描述语言”**

语料库

- 什么是语料库？
  - 语料库是语言材料的集合
- 语料库的特点
  - 真实语言环境中出现过的语言材料
  - 以电子计算机为载体
  - 经过一定的分析、加工和处理

语料库的类型

- 按来源分类
  - – 口语语料库
  - – 书面语语料库
- 按语言分类
  - – 单语语料库
  - – 双语语料库

语料库的类型

- 按加工方式分
- – 单语
  - 原始语料库
  - 切分标注语料库
  - 句法树库
  - 语义标注语料库
  - . ……
- – 双语
  - 篇章对齐语料库
  - 句子对齐语料库
  - 词语对齐语料库
  - 结构对齐语料库
  - . ……

语料库研究的历史

- 第一代（1970－80年代）
  - – 百万词级
  - – 以语言研究为导向
- 第二代（1980－90年代）
  - – 千万词级
  - – 词典编纂－ 应用导向
- 第三代（1990年代－ ）
  - – 超大规模（上亿词级）
  - – 标准编码体系
  - – 深度标注/多语种
  - – NLP应用
- 第四代（？ ）– 互联网作为语料库

第一代语料库例

- **LOB语料库**
  - – 始建于1970年代初
  - – 由英国Lancaster大学著名语言学家Geoffrey Leech倡议
  - – 挪威Oslo大学Stig Johansson主持完成
  - – 安装在挪威Bergen大学挪威人文科学计算中心
  - – 规模与Brown语料库相当
  - – 主要代表当代英国英语

第二代语料库例

- **COBUILD语料库**
  - **– 建于1980年代**
  - **– 以词典编撰为应用背景**
  - **– 有英国Birminghan大学与Collins出版社合作完成**
  - **– 规模达2000万词次**
  - **– 基于该语料库出版的Collins Cobuild词典（1987）受到了广泛的好评**

第三代语料库例

- **ACL/DCI语料库**
  - – 收集语料范围广泛
    - 华尔街日报
    - Collins英语词典
    - Brown语料库
    - PennTreeBank
    - 一些双语或多语文本等
  - – 既有已标注的语料，也有未标注语料
  - – 制定了语料库文件的格式标注
    - 采用统一的SGML标注语言
    - 语料标注依照TEI(Text Encoding Initiative)标准

总结：语料库的三个方面

语料本身	规模	百万/千万/亿万/…
	领域	政治/经济/体育/心理学/…
	体裁	文学/应用文/新闻/…
	语体	书面语/口语
	语种	单语/双语/多语
语料加工	数据形式	Text/HTML/数据库…
	编码体系	TEI标注/自定义编码体系…
	加工层次	词性/句法/语义/篇章/…
		双语句子对齐/词对齐/…
	加工方式	自动/人机互助/人工
语料应用	应用领域	通用/词典编撰/机器翻译/…

语料库例1:《人民日报》语料库

- 北京大学、富士通公司、人民日报社共同开发
- 包含了《人民日报》1998年上半年全部文本（约1千7百万字）
- 完整的词语切分和词性标注信息
- 高准确率

[中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/w 并/c 按照/p “/w 一国两制/j ”/w 、/w “/w 港人治港/l ”/w 、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an 。/w

人民日报语料库部分标注集

Ag	形语素	形容词性语素
a	形容词	
ad	副形词	直接作状语的形容词
an	名形词	具有名词功能的形容词
b	区别词	
c	连词	
Dg	副语素	
d	副词	
e	叹词	
f	方位词	
g	语素	
h	前接成分	
i	成语	
j	简称略语	
l	习用语	

语料库例2: Chinese Penn Tree Bank

- 句法树标注结果：

```
(IP (NP-SBJ (PN 他))
    (VP (ADVP (AI) 还))
        (VP (VV 提出)
            (NP-OBJ (QP (CD 一)
                        (CLP (M 系列))
                    (NP (NP (ADJP (JJ 具体))
                            (NP (NN 措施)))
                        (CC 和)
                        (NP (NN 政策)
                            (NN 要点))))))
    (PU 。))
```

语料库的编码体系

- SGML（标准置标语言）
  - http://www.w3.org/MarkUp/SGML/
- XML（可扩展的置标语言）
  - http://www.w3.org/TR/REC-xml
- TEI（文档编码计划）
  - http://www.tei-c.org/
- CES（语料库编码标准）
  - http://www.tei-c.org/Applications/index-co02.html

2. 机器翻译

机器翻译

- 机器翻译的历史
- 机器翻译的分类
- 机器翻译的范式
- 机器翻译的基本策略
- 基于规则的机器翻译方法
- 基于实例的机器翻译方法
- 统计机器翻译(SMT)
- 基于深度学习的机器翻译

机器翻译的定义(MT)

- 机器翻译(machine translation)是用计算机实现不同语言之间的翻译
  - Source language → target language

机器翻译的历史

- 机器翻译经历了以下几个阶段：
  - 萌芽期
  - 草创期
  - 萧条期
  - 复苏期
  - 繁荣期

萌芽期

- 上世纪三十年代，亚美尼亚裔的法国工程师阿尔楚尼（**G.B. Artsouni**）提出了用机器来进行语言翻译的想法，并在**1933**年**7**月**22**日获得了一项“翻译机”的专利，当时被称为“机械脑”。

- 阿尔楚尼认为这种机械脑尤其适合作机器词典，在宽纸带上面，每一行记录了源语言的一个词项以及这个词项在多种目标语言中的对应词项，在另外一条纸带上对应的每个词项处，记录着相应的代码，这些代码以打孔来表示。

草创期

- 1949年，美国洛克菲勒基金会副总裁韦弗发表了一份以《翻译》为题的备忘录，正式提出机器翻译问题。

- 备忘录中他提出各种语言有许多共同的特征，认为当把语言A翻译为语言B时，可以认为是从语言A出发，经过某一“通用语言”或“中间语言”转换为语言B，这种通用语言可以假定是全人类共同的。

- 1954年，美国乔治敦大学在IBM公司的协同下，用IBM-701计算机进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语句子。接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

萧条期

- 1964年，美国科学院成立语言自动处理谘询委员会(ALPAC)，于1966年11月公布了一个题为《语言与机器》的报告，简称ALPAC报告，宣称：在目前给机器翻译以大力支持还没有多少理由，报告还指出，机器翻译研究遇到了难以克服的“语义障碍”。

- 在ALPAC报告的影响下，许多国家的机器翻译研究进入低谷， 许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内机器翻译出现空前萧条的局面。

复苏期

- 出现实用的机器翻译系统，典型代表：
  - 法国格勒诺布尔理科医科大学应用数学研究所自动翻译中心研制出的俄法机器翻译系统，并接近实用水平。
  - 美国斯坦福大学威尔克斯提出了“优选语义学”（preference semantics），在此基础上设计了英法机器翻译系统，系统能有效解决仅用句法分析方法难以解决的歧义、代词所指等困难问题。

繁荣期

- 1976年--现在，机器翻译进入繁荣期;
- 特点是机器翻译研究走向了实用化和商品化。

机器翻译的分类

- 理想的机器翻译
  - 全自动高质量
- 按人机关系分类
  - 全自动机器翻译
  - 人助机译
  - 机助人译

机器翻译的分类

- 按应用方式分类
  - 信息分发型：
    - 要求高质量，不要求实时
    - 采用人机互助或者受限领域、受限语言等方式
- 提高翻译质量
  - 信息吸收型：
    - 不要求高质量，要求方便、实时
    - 翻译浏览器、便携式翻译设备、……

机器翻译的分类

- 按应用方式分类（续）
  - 信息交流型
    - 不要求高质量，通常要求实时，语言随意性较大
    - 语音翻译、网络聊天翻译、电子邮件翻译
  - 信息存取型
    - 将机器翻译嵌入到其他应用系统中
    - 跨语言检索、跨语言信息抽取、跨语言文摘、跨语言非文本数据库的检索……

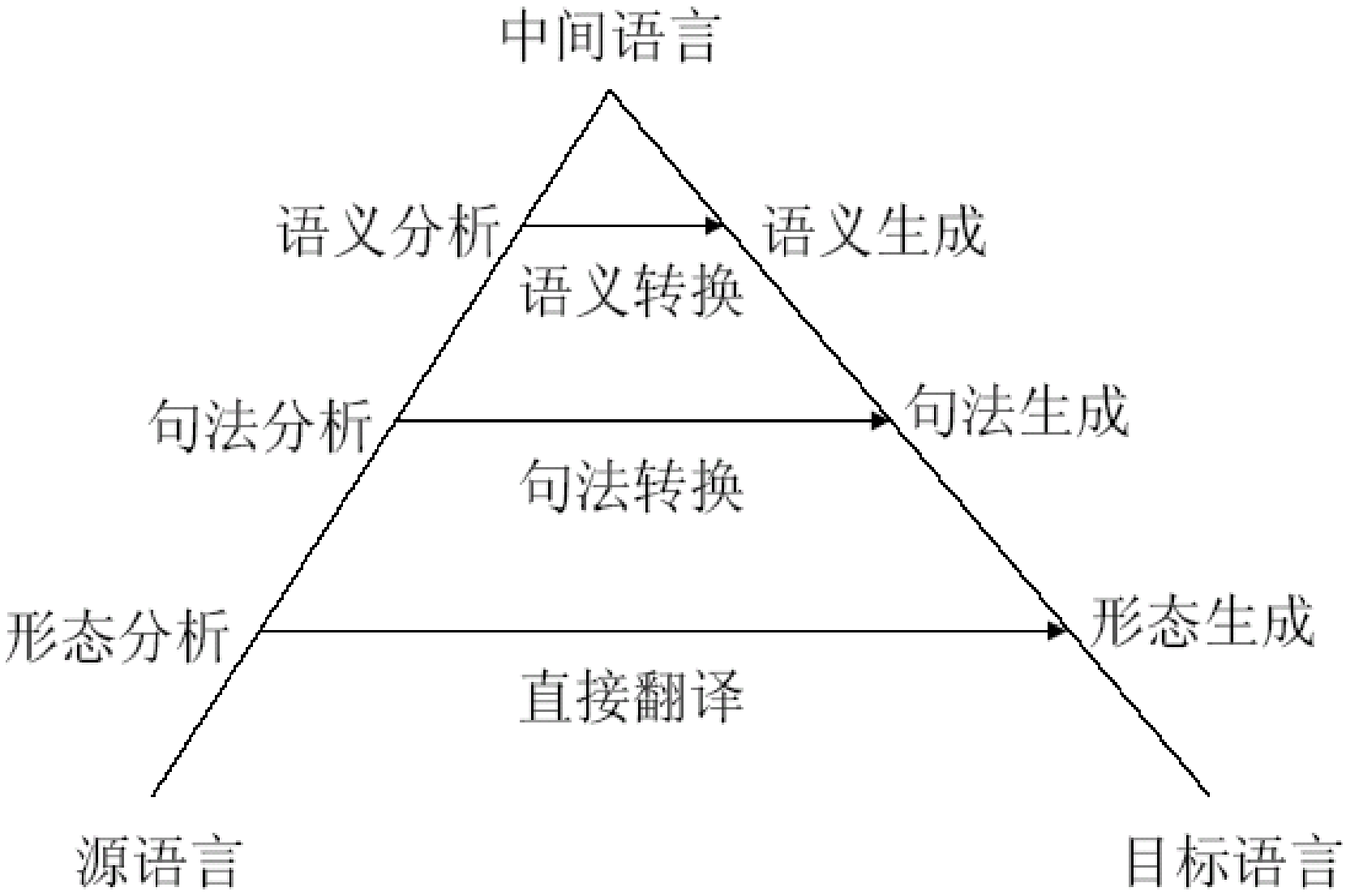
机器翻译系统-例1

- 口语翻译系统：
  - 1989年，日本ATR研制了SL-TRANS口语翻译系统。
  - 90年代，日本学者北野采用基于实例的方法进行语音翻译实验，证明了毫秒级的实时口语语音翻译是可实现的。
  - **Verbmobil计划-建立非特定人的、面向会面安排交谈的口语语音翻译系统。**
  - **C-STAR计划，多个国家共同完成面向口语的机器翻译平台。**

机器翻译系统-例2

- 互连网与机器翻译:
  - 词典容量大;
  - 翻译速度快，译文具有可读性;
  - 译文质量较粗，不要求做到译文的“达”和“雅”。

传统机器翻译范式



经典机器翻译方法

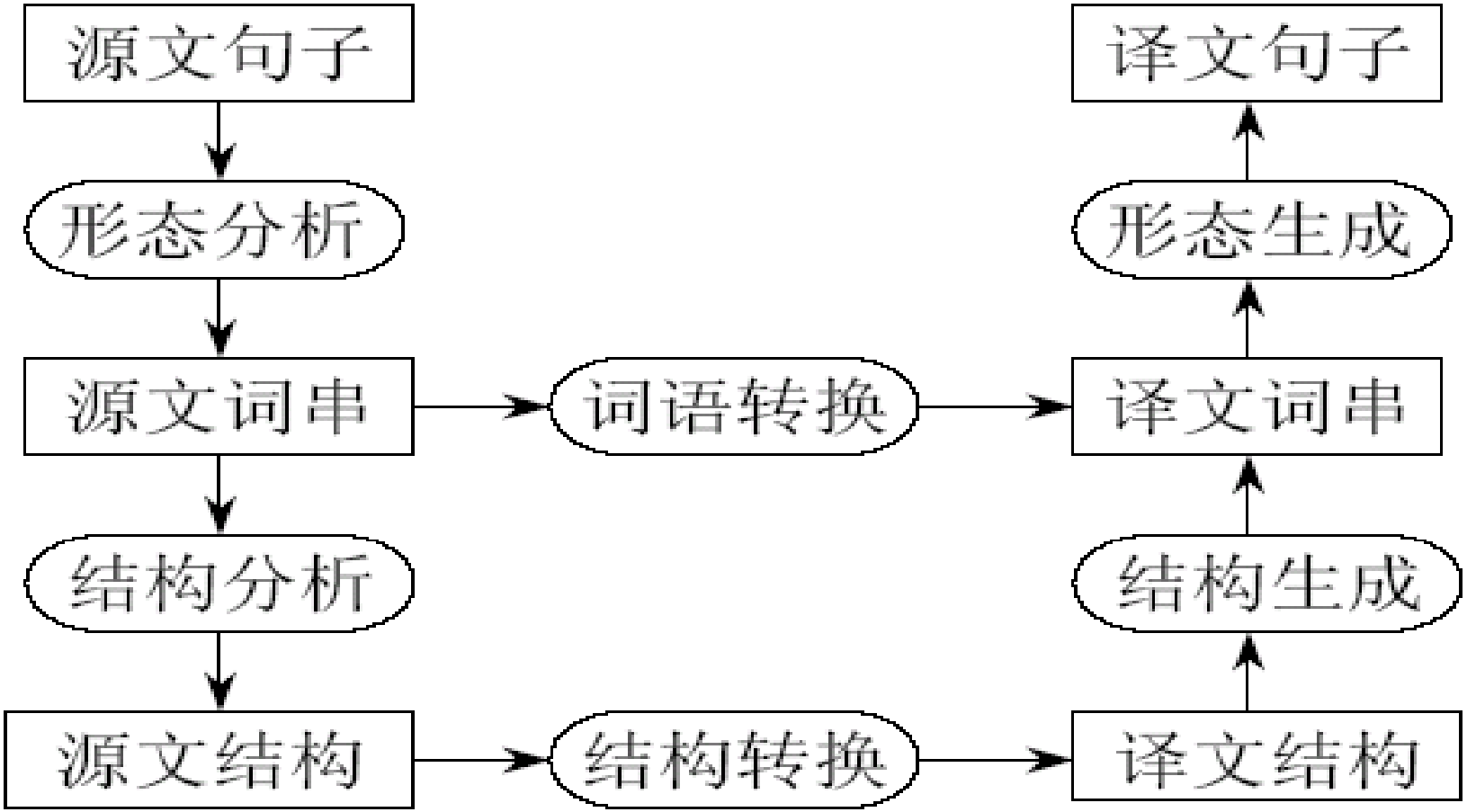
(1)直接翻译方法

- 通过词语翻译、插入、删除和局部的词序调整来实现翻译，不进行深层次的句法和语义的分析，但可以采用一些统计方法对词语和词类序列进行分析

(2)转换方法

- 翻译分为“分析”“转换”“生成”三个阶段：
  - 分析：源语言句子→源语言深层结构
  - 转换：源语言深层结构→目标语言深层结构
  - 生成：目标语言深层结构→目标语言句子

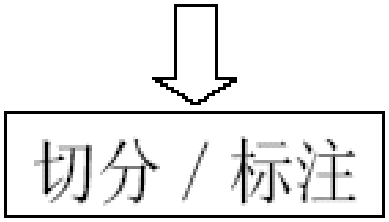
转换方法例



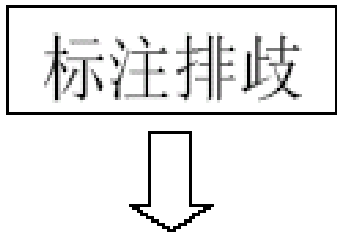
基于转换方法的翻译流程

词法分析

她把一束花放在桌上。⇨ She put a bunch of flowers on the table.



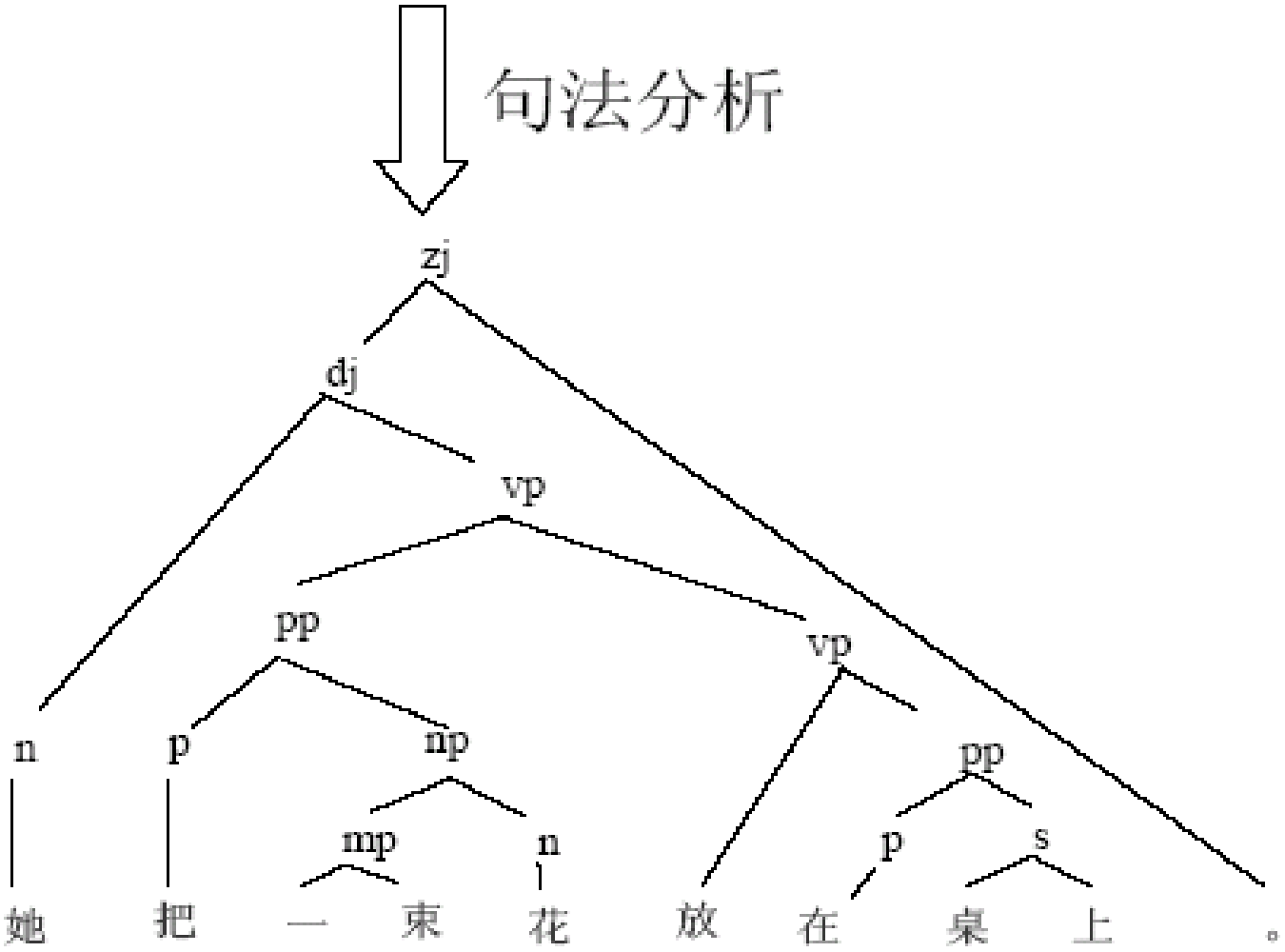
她/r 把/p-q-v-n 一/m-d 束/q 花/n-v-a 放/v 在/p-d-v 桌/n 上/f-v 。/w



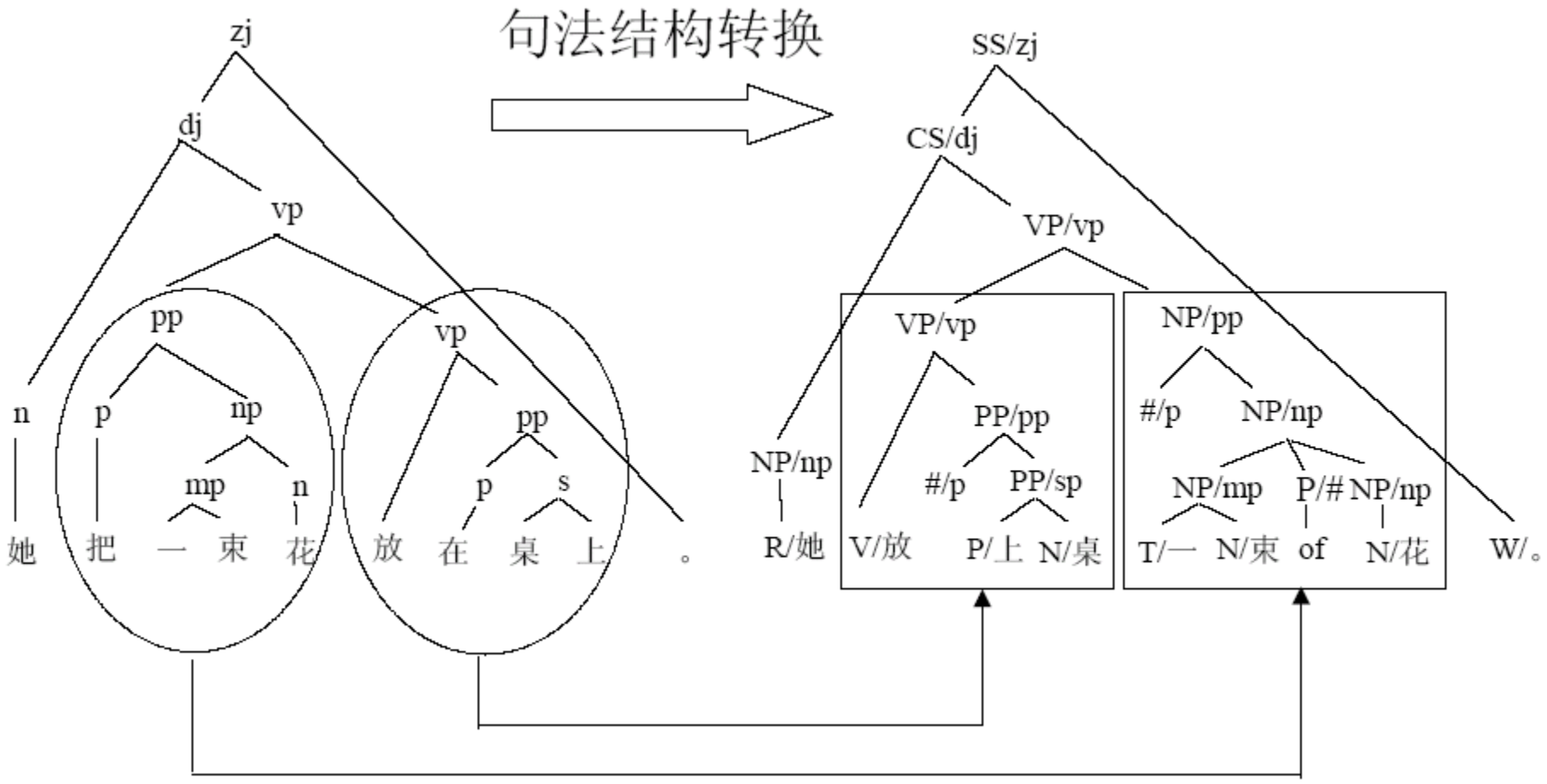
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。/w

句法分析

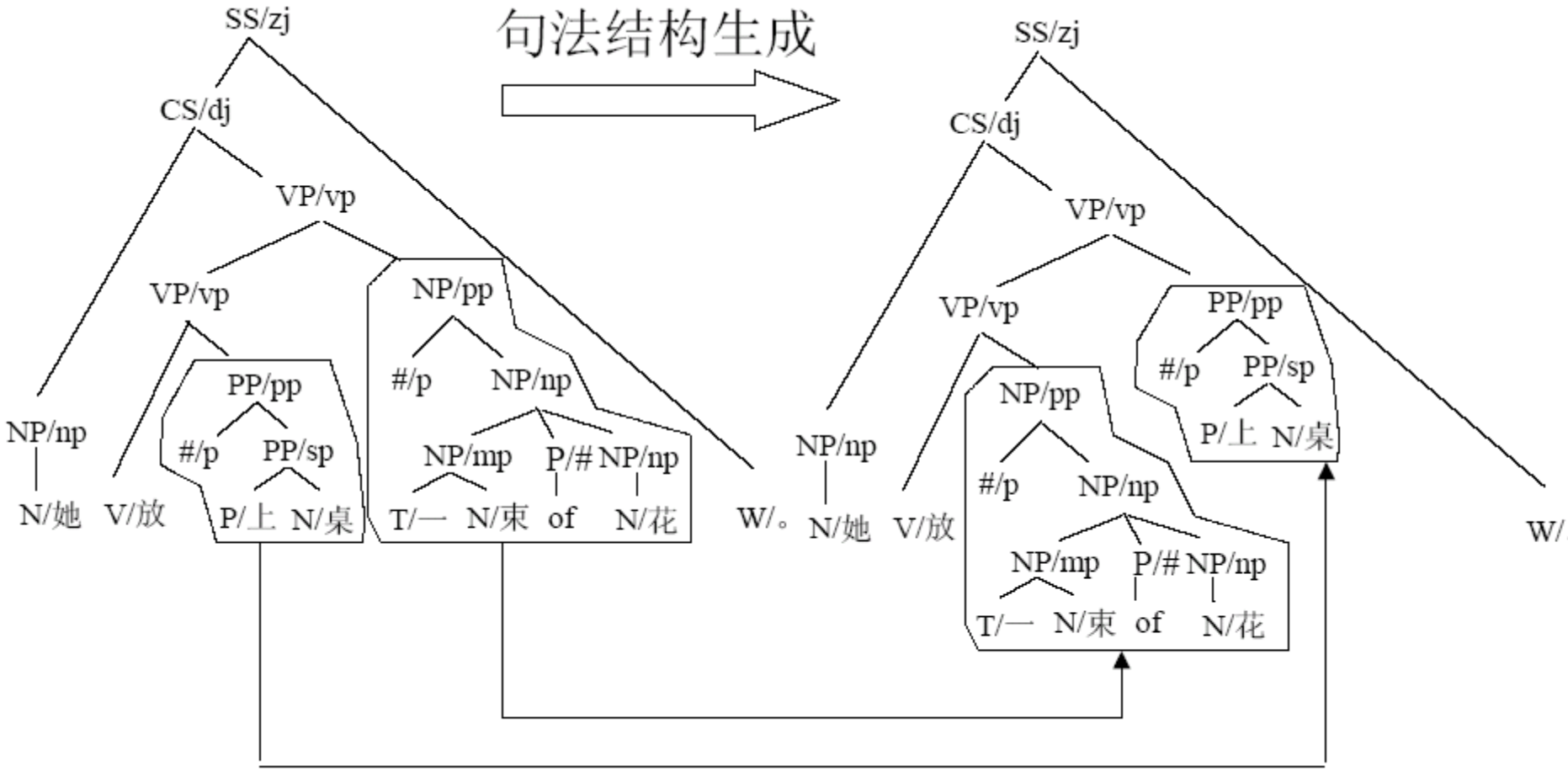
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。/w



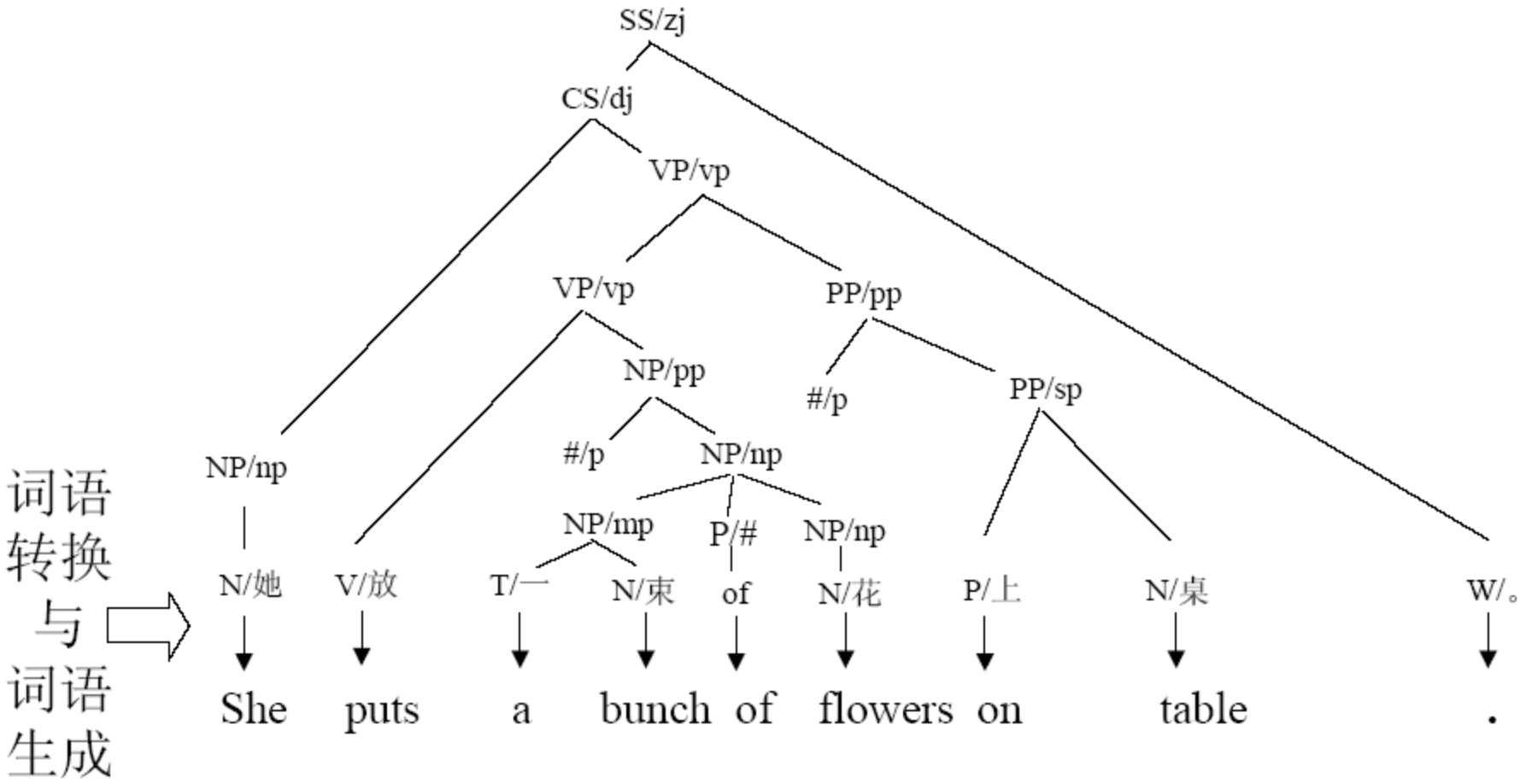
句法结构转换


句法结构转换
zj
dj
vp
pp
p
np
mp
n
她
把
一
束
花
vp
pp
p
s
放
在
桌
上
。
SS/zj
CS/dj
VP/vp
NP/np
R/她
VP/vp
PP/pp
#/p
PP/sp
V/放
P/上
N/桌
NP/pp
#/p
NP/np
NP/mp
P/#
NP/np
T/一
N/束
of
N/花
W/。

目标语句法结构生成


句法结构生成
SS/zj
CS/dj
VP/vp
VP/vp
NP/pp
PP/pp
#/p
NP/np
#/p
PP/sp
NP/mp
P/#
NP/np
NP/np
N/她
V/放
P/上
N/桌
T/一
N/束
of
N/花
W/。
SS/zj
CS/dj
VP/vp
VP/vp
PP/pp
#/p
PP/sp
P/上
N/桌
NP/pp
#/p
NP/np
NP/mp
P/#
NP/np
NP/np
N/她
V/放
T/一
N/束
of
N/花
W/。

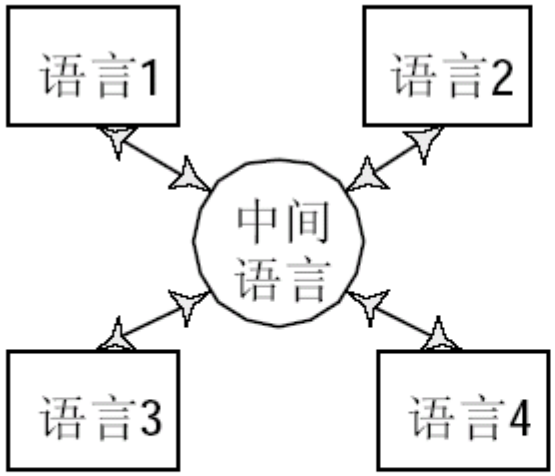
目标语词语转换


SS/zj
CS/dj
VP/vp
VP/vp
PP/pp
NP/pp
#/p
PP/sp
#/p
NP/np
NP/np
NP/mp
P/#
NP/np
N/她
V/放
T/一
N/束
of
N/花
P/上
N/桌
W/。
She
puts
a
bunch
of
flowers
on
table
.
词语转换与词语生成

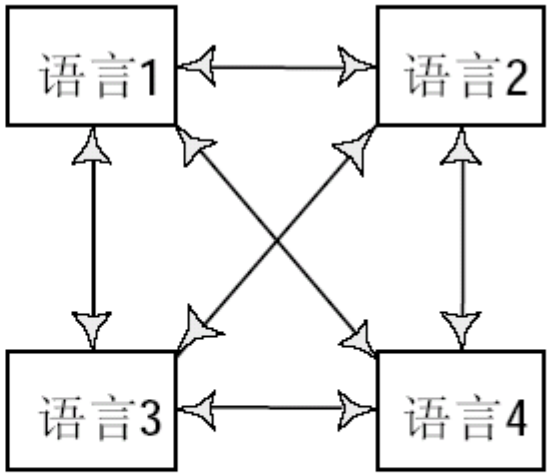
(3)中间语言方法

- 利用一种中间语言作为翻译的中介表示形式；
- 翻译过程分为“分析”和“生成”两个阶段
  - 分析：源语言→中间语言
  - 生成：中间语言→目标语言
  - 分析过程只与源语言有关，与目标语言无关
  - 生成过程只与目标语言有关，与源语言无关

(3)中间语言方法



中间语言方法



转换方法

(3)中间语言方法

- 中间语言的类型
  - 自然语言：如英语、汉语
  - 人工语言：如世界语
  - 某种知识表示形式：如语义网络
- 以某种知识表示形式作为中间语言的机器翻译方法有时也称为基于知识的机器翻译方法

机器翻译的基本策略

机器翻译的基本策略

- 基于规则的机器翻译方法
- 基于语料库的机器翻译方法
  - 基于实例的机器翻译方法
  - 基于翻译记忆的机器翻译方法
  - 基于统计的机器翻译方法
- 多引擎机器翻译方法
- 基于深度学习技术的机器翻译方法

(1)基于规则的方法

- 采用规则作为知识表示形式：
  - 重叠词规则
  - 切分规则
  - 标注规则
  - 句法分析规则
  - 语义分析规则
  - 结构转换规则（产生译文句法语义结构）
  - 词语转换规则（译词选择）
  - 结构生成规则（译文结构调整）
  - 词语生成规则（译文词形生成）

(1)基于规则的方法

- 优点
  - 直观，能够直接表达语言学家的知识
  - 规则的颗粒度具有很大的可伸缩性
    - 大颗粒度的规则具有很强的概括能力
    - 小颗粒度的规则具有精细的描述能力
  - 便于处理复杂的结构和进行深层次的理解， 如解决长距离依赖问题
  - 系统适应性强，不依赖于具体的训练语料

(1)基于规则的方法

- 缺点
  - 规则主观因素重，有时与客观事实有一定差距
  - 规则的覆盖性差，特别是细颗粒度的规则很难总结得比较全面
  - 规则之间的冲突没有好的解决办法（翘翘板现象）
  - 规则一般只局限于某一个具体的系统，规则库开发成本太高
  - 规则库的调试极其枯燥乏味

基于规则的方法—译词选择

```
$$ 开
    **{v} v $=[…]
        || $.主体=是,$.主体.语义类=植物
        → V<bloom> $=[…]
        || $.客体=是,$.客体.汉字=灯|机|器
        → V( !V<turn> D<on> ) $=[…]
        || $.客体=是,$.客体.语义类=交通工具
        => V<drive> $=[…]
        || OTHERWISE
        => V<open> $=[…]
```

基于规则的方法—结构转换

&& {mp7} mp->r !mp :: $.内部结构=组合定中,…

|| %mp.定语.内部结构=单词, %mp.定语.yx=一,%mp.量词子类=集体|种类|容量|时量|度量|成形

=> NP(T/r !NP/mp) %T.TNNUM=%NP.NNUM /*这一年*/

|| %mp.定语.内部结构=单词, ,%mp.定语.yx=一,%mp.量词子类=个体

=> T(T/r M<one>) /*这一个 哪一个*/

|| %r.yx=这|那, IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE

=> NP(T/r !M/mp) %T.TNNUM=PLUR,$.NNUM=PLUR /*这两张*/

=> NP(T/r !NP/mp) %T.TNNUM=PLUR,$.NNUM=PLUR

|| %r.yx=~这~那,IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE

=> NP(T/r !M/mp) $.NNUM=%M.NNUM

=> NP(T/r !NP/mp) %T.TNSUB=%NP.NSUBC,…

基于规则的方法—结构生成

```
## { NPMP1 } NP(T !NP(T !N))
        => NP(T/T !NP/NP(!N/N))
        /* this a kind => this kind */
## { NPATN1 } NP(AP(!A) !NP(T !N))
        => P(T/T !NP/NP(AP/AP(!A/A) !N/N))
        /* red this book => this red book */
```

(2)基于语料库的机器翻译方法

- 优点
  - 使用语料库作为翻译知识来源，无需人工编写规则， 系统开发成本低，速度快
  - 从语料库中学习到的知识比较客观
  - 从语料库中学习到的知识覆盖性比较好缺点
  - 系统性能依赖于语料库
  - 数据稀疏问题严重
  - 语料库中不容易活动大颗粒度的高概括性知识

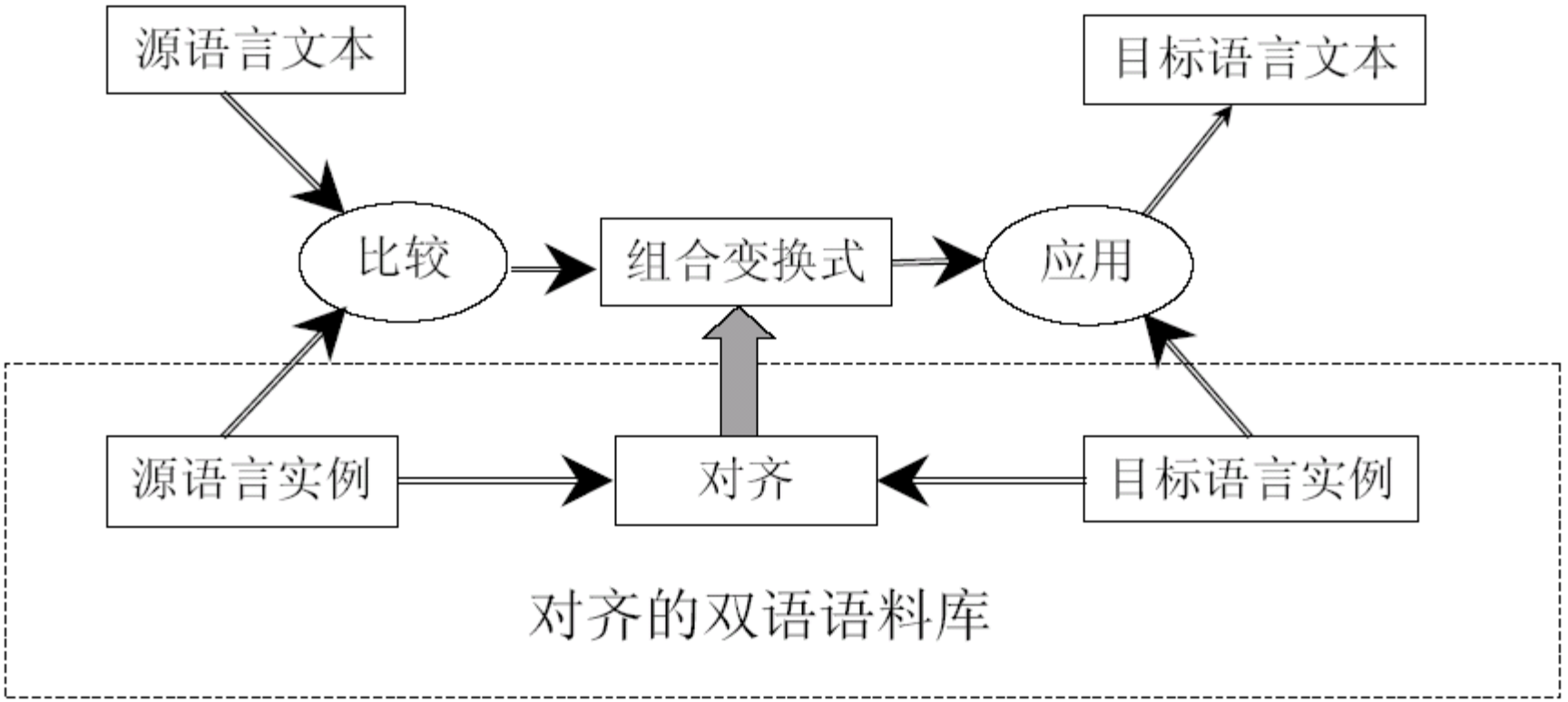
(3)基于实例的机器翻译

- 长尾真(Makoto Nagao)在1984年指出，人类并不通过做深层的语言学分析来进行翻译，人类的翻译过程是：首先把输入的句子正确地分解为一些短语碎片，接着把这些短语碎片翻译成其它语言的短语碎片，最后再把这些短语碎片构成完整的句子， 每个短语碎片的翻译是通过类比的原则来实现的。
  - 如果我们给出一些英语句子的实例以及相对应的日语句子，机器翻译系统来识别和比较这些实例及其译文的相似之处和相差之处，从而挑选出正确的译文。
  - 因此，我们应该在计算机中存储一些实例，并建立由给定的句子找寻类似例句的机制，这是一种由实例引导推理的机器翻译方法。

(3)基于实例的机器翻译

- 优点
  - 直接使用对齐的语料库作为知识表示形式， 知识库的扩充非常简单
  - 不需要进行深层次的语言分析，也可以产生高质量的译文
- 缺点
  - 覆盖率低，实用的系统需要的语料库规模极大（百万句对以上）

基于实例的机器翻译系统结构



基于实例的机器翻译一举例

要翻译句子：

(E1) He bought a book on physics.

在语料库中查到相似英语句子及其汉语译文是：

(E2) He wrote a book on history.

(C2) 他写了一本关于历史的书。

比较(E1)和(E2)两个句子，我们得到变换式：

(T1) replace(wrote, bought) and replace(history, physics)

将这个变换式中的单词都换成汉语就变成：

(T2) replace(写,买) and replace(历史,物理)

将(T2)作用于(C2)

(C1)他买了一本关于物理学的书。

基于实例的机器翻译系统例

- 日本京都大学长尾真和佐藤研制的**MBT1**和**MBT2**系统
- 美国卡内基-梅隆大学研制的**PANGLOSS**系统；
- 日本口语翻译通信研究实验室**ATR**研制的**ETOC**和**EBMT**系统；
- 我国清华大学研制的基于实例的日汉机器翻译系统；

(4)基于翻译记忆的机器翻译方法

- 翻译记忆的基本思想：
  - 把已经翻译过的句子保存起来，翻译新句子时直接到语料库中去查找，如果发现相同的句子，直接输出译文，否则交给人去翻译，但可以提供相似的句子的参考译文。
- 翻译记忆的主要关键技术：
  - 计算待翻译内容与记忆库中实例之间的相似度，以获取最相似的实例；
  - 参考最相似实例的译文构造待译内容的译文；
  - 根据翻译记忆全过程的需要，设计翻译记忆的相关记忆库。

(4)基于翻译记忆的机器翻译方法

- 翻译记忆方法主要被应用于计算机辅助翻译软件中，其优缺点：
  - 翻译质量有保证
  - 随着使用时间匹配成功率逐步提高
  - 特别适用于重复率高的文本翻译，例如公司的产品说明书的新版本翻译
  - 与语言无关，适用于各种语言对
  - 缺点是使用初期匹配成功率不高
- 相关产品：
  - **TRADOS**
  - 雅信CAT

(5)基于深度学习技术的机器翻译

- 基于深度学习技术的机器翻译利用深度学习技术改进了统计机器翻译中的一些关键问题和模块，并取得了较大的提升。

- 目前，**GOOGLE**上线的机器翻译系统已经由原来的基于统计的改为基于深度学习技术的机器翻译系统；微软和**facebook**等也相继使用相关的深度学习技术改进机器翻译。

机器翻译现状(截止2017.5)

- **Google翻译**
- **百度翻译**
- **微软翻译**
- **Facebook翻译**
  - 刚到五月，你是不是已经感受到了炎炎夏日的威力？据中央气象台消息，本轮高温过程将从5月17日持续到19日。其中，18日的高温范围最大、强度最强。

google

刚到五月，你是不是已经感受到了炎炎夏日的威力？据中央气象台消息，本轮高温过程将从5月17日持续到19日。其中，18日的高温范围最大、强度最强

- **Just arrived in May, you are not already feel the power of the summer heat? According to the Central Meteorological Observatory news, the current high-temperature process will be from May 17 to 19 days. Among them, the 18 day high temperature range, the strongest intensity.**

bing

刚到五月，你是不是已经感受到了炎炎夏日的威力？据中央气象台消息，本轮高温过程将从5月17日持续到19日。其中，18日的高温范围最大、强度最强

- Did you feel the power of the scorching summer just by May? According to the Central Meteorological Observatory, the high temperature process will continue from May 17 to 19th. Among them, 18th of the highest temperature range, the strongest strength.

baidu

刚到五月，你是不是已经感受到了炎炎夏日的威力？据中央气象台消息，本轮高温过程将从5月17日持续到19日。其中，18日的高温范围最大、强度最强

- **In May, have you felt the power of the scorching summer? According to the Central Meteorological Observatory news, this round of high temperature process will continue from May 17th to 19. Among them, the 18 day high temperature range, the strongest intensity.**

3. 自动问答

相关自动问答系统例

- 回答依然是网页的形式：
  - 美国AskJeeves 公司的检索系统
    - www.ask.com
    - 香港科技大学
    - http://www.weniwen.com
- 直接给答案的形式：
  - 麻省理工(MIT)开发的问答系统Start
    - http://www.ai.mit.edu/projects/infolab/
  - AnswerBus
    - http://misshoover.si.umich.edu/~zzheng/qa-new

what's the capital of China?

Ads related to: **what's the capital of China?**

Capital China - Incredibly Great Prices - tripadvisor.com
www.tripadvisor.com/ ▼
TripAdvisor® checks prices—on up to 200 sites—to help you find the best deals.
Easy price comparison · Millions of hotel reviews · Candid traveler photos

Web Results

Source

The Capital of China is Beijing.

Largest City: Shanghai (17,900,000 people)

Sources:

1. CIA World Factbook - https://www.cia.gov/library/publications/the-wo...
2. Wikipedia - http://en.wikipedia.org/wiki/List_of_national_c...

See Also: Official Site · BBC Profile · Encyclopedia

Search For: Government · Economy · People · Largest City

Capital of: Afghanistan ▼ Go

China - Wikipedia
en.wikipedia.org/wiki/China
China, officially the People's Republic of China (PRC), is a unitary sovereign state in East Asia ..... By the time of the Warring States period of the 5th–3rd centuries BCE, there were seven powerful sovereign states in what is now China, each with its .... In the early years of the Ming Dynasty, China's capital was moved from ...

START
Natural Language Question Answering System

What's the capital of China? Ask Question

===> What's the capital of China

China

Capital:
name: Beijing
geographic coordinates: 39 55 N, 116 23 E
time difference: UTC+8 (13 hours ahead of Washington, DC during Standard Time)
note: despite its size, all of China falls within one time zone; many people in Xinjiang Province observe an unofficial "Xinjiang time zone" of UTC+6, two hours behind Beijing

Source: The World Factbook

Beijing is the capital of China.

Source: START KB

自动问答

- **60年代：人工智能研究刚刚开始时，人们提出让计算机用自然语言来回答人们的问题；**

- **80年代：研究者们纷纷开始研究自然语言问答系统，由于当时条件限制，实验都是在受限领域甚至是固定段落上进行的，所以问答系统一直被限制在特殊领域的专家系统.**

自动问答

- **90年代后：自动问答备受关注，微软、IBM 等大公司投入研究问答系统；著名自动问答系统有Start、AskJeeves、AnswerBus，MULDER、LAMP等，其中Start是麻省理工学院开发的问答系统。**

问答系统例

- **Start系统是第一个面向互联网的自然语言问答系统，它能够回答一些有关地理、历史、文化、科技、娱乐等方面的简单问题**

- **AskJeeves是美国AskJeeves公司的检索系统，返回的结果是网页，不是问题的直接答案。**

- **AnswerBus是多语种自动问答系统，可以回答英语、法语、西班牙语、德语等很多语种的问题，但返回的是可能包含答案的句子和相关联的URL。**

问答系统例

- 华盛顿大学的**MULDER**是第一个以网络作为其知识库的自动问答系统，它将检索到的网络文档下载到本地，并且对这些文档进行详细的语法分析从中抽取答案

- 新加坡国立大学的**LAMP**是和**MULDER**相似的系统，仅分析搜索引擎返回的网页片断信息，采用传统的向量空间模型**(VSM)**的改进作为抽取答案的方法。

百度

Bai du 百度 中国的首都是什么？

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约15,100,000个 搜索工具

中华人民共和国首都：

北京市

北京，简称"京"，中华人民共和国首都、直辖市、国家中心城市、超大城市，全国政治中心、文化中心、国际交往中心、科技创新中心，是中国共产党中央委员会、中华人民共和... 详情>>

来自百度百科 | 报错

中国的首都是什么?还有各国的首都吗?_百度知道

1个回答 - 提问时间: 2015年02月26日

[专业] 答案:中国的首都是北京。各国首都:【亚洲】·中国:北京·韩国:首尔·日本:东京·泰国:曼谷·马来西亚:吉隆坡·越南:河内·朝鲜:平壤·印度:新德里·文莱:斯里...

zhidao.baidu.com/link?... - 评价

我国的首都是什么? 2个回答 2013-12-27

更多知道相关问题>>

问答系统的核心

- 问题理解
- 信息检索
- 答案抽取

问答系统的范式

- 基于IR的方法：
  - TREC; IBM Watson; Google
- 基于知识和基于IR的混合方法：
  - IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi

1. 基于IR的方法

答案在 IR中？


Bai du 百度
广外位于哪里？
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约280,000个
搜索工具
广外在哪_百度知道
3个回答 - 提问时间: 2011年11月06日
最佳答案: 你问哪个校区? 本部=北校区=白云山校区,公交车坐到 外语学院(白云山西门)站/黄石东路口站,下车请问路,大约还有3分钟脚程 地铁坐到 白云公园/萧岗/...
zhidao.baidu.com/link?... - 评价
广外在哪里? 2个回答 2006-11-04
更多知道相关问题>>
广外_百度地图
在以下城市有结果，请您选择：
北京市(766) 广州市(957) 深圳市(3) 南宁市(36) 南昌市(22) 佛山市(7)
map.baidu.com
广东外语外贸大学校本部在哪?_百度知道
2个回答 - 最新回答: 2012年10月15日 - 12人觉得有用
这个是广外最最根本、中心的校区了,因为大学城那边是新建的,北校区是开始就有...2013-10-15 广东外语外贸大学南校区在哪里?详细点! 21 2012-09-04 广东...
更多关于广外位于哪里?的问题>>
zhidao.baidu.com/link?... - 百度快照 - 评价
广外在哪里?_百度知道

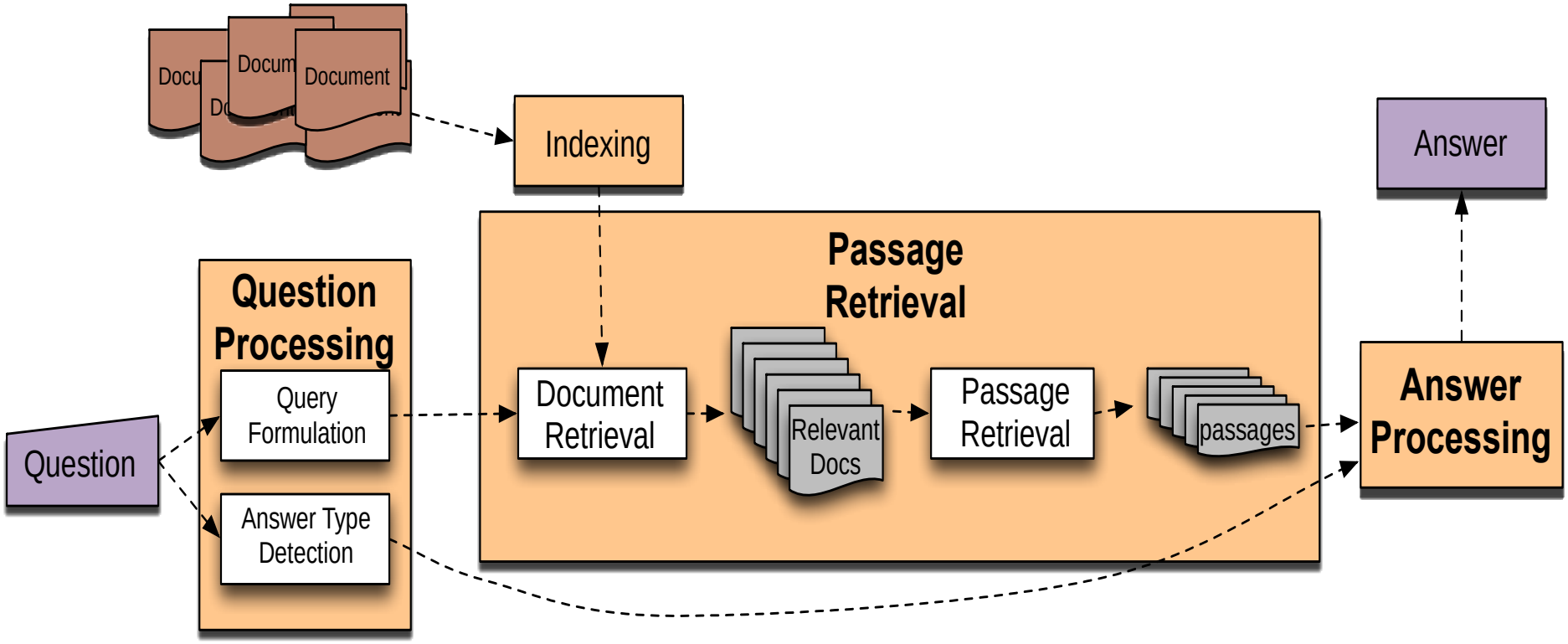
答案在 IR中？


Bai du 百度
功夫熊猫什么时候播出？
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约1,130,000个
搜索工具
问 功夫熊猫什么时候播出？：
2016电影功夫熊猫什么时候上映_百度知道
《功夫熊猫3》是《功夫熊猫》系列的第三部电影,由余仁英执导,杰克·布莱克、凯特·哈德森、布莱恩·科兰斯顿原版配音主演,黄磊、杨幂、成龙、周杰伦、朱珠...
来自百度知道 | 3个回答 | 2016-01-20
功夫熊猫一是什么时候上映的?_百度知道
1映时间:2008年5月 功夫熊猫22011年5月26日下午,3D版《功夫熊猫2》在北京UME影城举行首映看片,熊猫阿宝带着盖世五侠,再次以它憨态十足的搞笑功力而来.这一次阿...
来自百度知道 | 4个回答 | 2013-11-23
功夫熊猫上映和下架时间_百度知道
2016年1月29日中美同步正式上映,由于被认定为中美合拍片,并不占用引进片名额,不知道会不会考虑延期下映.
来自百度知道 | 2个回答 | 2016-01-24
功夫熊猫什么时候上映_功夫熊猫上映时间_百田电影大全
功夫熊猫什么时候上映?美国动画喜剧动作冒险电影《功夫熊猫》上映时间为2008年06月20日,

答案在 IR中？


Bai du 百度
我国的四大发明是？
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约2,090,000个
搜索工具
问 我国的四大发明是什么_百度知道
答 我国的四大发明是 司南 造纸术 活字印刷术 火药
来自百度知道 | 报错
我国四大发明有哪些 5个回答 2011-05-26
中国古代四大发明是什么 2个回答 2013-11-28
更多相关问答>>
我国古代四大发明分别在哪个时期? - 爱问知识人
2个回答 - 最新回答: 2010年2月1日
最佳答案: 1是造纸术，2是印刷术，3是指南针，4是火.药，造纸术是汉代蔡伦发明的，印刷术，是宋代毕升发明的，，指南针，是上古的黄帝所发明 的火.药，是...
iask.sina.com.cn/b/166... - 百度快照 - 2041条评价
【你知道我国四大发明是什么】百度作业帮
1个回答 - 提问时间: 2014年11月19日
最佳答案: 一、造纸术造纸术在东汉的时候就已由蔡伦发明,蔡伦发明以来只就可以替代笨重的竹简在汉朝的谷树皮的谷纸的基础上,晋代出现了以藤皮做出的藤纸,造纸术在...
https://www.zybang.com/questio... V2 - 百度快照 - 5031条评价

基于IR的自动问答基本范式



基于IR的自动问答的基本步骤

- 问题预处理
  - 问题类型检测
  - 问题的关键词
- 篇章检索
  - 依据问题关键词搜索相关文档；
  - 对检索到的文档通过分段、分句等操作后，重新排序
- 答案抽取
  - 在检索到的文档中抽取相应的候选答案
  - 对这些候选答案评分排序

基于知识的方法 例Siri

- 对查询使用语义进行表示，包括：**Build a semantic representation of the query**
  - 时间、日期、地点、实体、数字等
- 将这些语义集合映射到结构化的数据或知识库
  - 地理数据库
  - 基于维基百科本体库
- 基于这些知识库的输入推理得到输出答案

混合方法 (IBM Watson)

- 建立查询的浅层语义表示
- 使用IR方法生成候选答案集合
  - 将这些候选答案扩展到本体库中
- 依据丰富的知识库资源对每个候选答案打分排序

自动问答系统基本设计方法

- 问题分类
- 问题关键字抽取
- 知识检索
- 答案抽取