认知基础
Cognitive Foundation
第六章

认知语言学
Cognitive Linguistics

史忠植

中国科学院计算技术研究所
http://www.intsci.ac.cn/

内容提要

- 概述
- 语言理解认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

语 言

- 语言是人类最重要的交际工具，是人们进行沟通的主要表达方式。人们借助语言保存和传递人类文明的成果。
- 语言是民族的重要特征之一。一般来说，各个民族都有自己的语言。汉语、法语、俄语、西班牙语、阿拉伯语、英语是世界上的主要语言，也是联合国的工作语言。
- 语言是由词汇按一定的语法所构成的复杂的符号系统，它包括语音系统、词汇系统和语法系统。

语　言

- 自然语言：人类交流的语言，口语、书面语、手语、旗语等
- 人造语言：机器语言，包括C++，BASIC等
  世界语
- 到目前为止的人类知识有80%以上使用自然语言文字记载下来的。但将来，可能用计算机语言形式记载的知识将会越来越多。因此说，语言信息处理技术和每年所处理的信息总量已成为衡量一个国家现代化水平的重要标志之一。
- 相比较人工智能其它领域，自然语言理解是难度大，进展小的。至今为止未能达到很高的水平。

语言处理

- 语言处理过程是指对信息的接收、存储、转化、传送和发布等
- 语言分级：字级处理、概念处理和智能处理
- 语言处理类别：

-计算语言学（Computational Linguistics）有时也叫,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）。

-认知语言学（Cognitive Linguistics）

计算语言学

- 计算语言学（Computational Linguistics）是通过建立形式化的数学模型，来分析、处理自然语言，并在计算机上用程序来实现分析和处理的过程，从而达到以机器来模拟人的部分乃至全部语言能力的目的。
- 自然语言理解的研究大体上经历了三个 时期
  - 萌芽时期
  - 发展时期
    - 早期：20世纪60年代以关键词匹配为主流
    - 中期：20世纪70年代以句法－语义分析为主流
    - 近期：20世纪80年代以来走向实用化和工程化
  - 大规模真实文本处理时期

认知语言学

- 认知语言学是认知心理学和语言学相结合的一个交叉边缘学科。它一方面从人的认知，即人们认识客观世界的方式的角度观察研究语言，另一方面，认知语言学把语言看作一种认知活动，认为语言是认知对世界经验进行组织的结果，是认知的重要组成部分。通过对语言现象的规则和普遍性的观察，分析语言所反映出的认知取向，从语言的各个层面探讨认知和语言的关系及其性质，说明语言是认知的产物，探讨人类的认知能力及其发展的规律性和共同性。

认知语言学

- 认知语言学形成与发展的直接动力源于语言学本身，最早是在语用学和生成语义学派的理论中认识到认知在语言中的作用。语用学研究一定语境中的语言，将人的认知体系看作语境的构成因素之一，这样就把对人的认知体系的研究纳入了语言研究之中。生成语义学对认知语言学产生的贡献在于，它认为语义是句法生成的基础，语义不能独立于人的认知。这就使得语言研究走上了认知语言学的新路。

认知语言学的创建

- 20世纪80年代，认知语言学取得了大的发展。1980年出版的莱考夫和约翰逊所著的《我们赖以生存的隐喻》一书以大量的语言事实论证了语言与隐喻认知结构的密切相关性。
- 1987年美国出版了标志认知语言学形成的三部研究专著，即约翰逊的《心中之身：意义、想象和理解的物质基础》、兰盖克的《认知语法基础》（第一卷）、莱考夫的《范畴》。
- 1989年在德国杜伊斯堡召开的第一届认知语言学学术会议标志着认知语言学的正式诞生，会上成立了国际认知语言学协会（ICLA），宣布发行《认知语言学》杂志，出版认知语言学研究专著。

认知语言学的基本研究框架

（1）语言研究必须同人的概念形成过程的研究联系起来。

（2）词义的确立必须参照百科全书般的概念内容和人对这一内容的解释(construal)。

（3）概念形成根植于普遍的躯体经验(bodily experience)，特别是空间经验，这一经验制约了人对心理世界的隐喻性建构。

（4）语言的方方面面都包含着范畴化，并以广义的原型理论为基础。

（5）认知语言学并不把语言现象区分为音位、形态、词汇、句法和语用等不同的层次，而是寻求对语言现象统一的解释。

认知语言学研究的主要内容

1、对自然语言的产生和理解的过程的研究。这是认知语言学最早的研究领域，研究的任务是力求对文本的理解和生成形成一定的模式，一些计算机程序方面的专家参与了大量的研究工作。这一领域研究取得突出成绩的语言学家首推契夫，正是他率先提出了现行知觉、激活等范畴性理论，由此派生出一系列认知语言学概念。

认知语言学研究的主要内容

2、对语言范畴化诸原则的研究。

- 原型是物体范畴最好、最典型的成员，所有其他成员也均具有不同程度的典型性。比如说，在英语的世界图景中，鸟的原型为画眉鸟；而对于母语为俄语的人而言则是麻雀；麻雀在中国人的认知意义中也具有典型意义。
- 根据莱考夫的论述，认知模式据其结构原则的不同分为四种：命题模式、意象图示模式、隐喻模式和转喻模式。
- 范畴化、原型、基本范畴和上下位概念等术语的提出对认知语言学的发展具有极其重要的意义。

认知语言学研究的主要内容

3、对概念结构及语言对应类型的研究。这方面的研究最初始于对人工智能的研究。

4、对认知语义高级范畴的研究。塔尔密是这方面研究的佼佼者。他尝试建立一套有序的形式构造层级范畴，并以此使自然语言实现对现实事物进行概念化结构操作。这些范畴如认知状态、布局结构、注意力分配等，每一范畴均具有自己复杂的结构。

认知语言学研究的主要内容

5、对语言中的空间关系及运动的概念化类型的研究。这是认知语言学中一个十分重要的研究领域，是诸多重要问题的结点。莱考夫提出了“意象图示”的重要概念并深入研究了诸如容器图示、部分- 整体图示、起点- 路径- 终点图示、中心- 边缘图示、连接图示等重要的图示类型。

6、对认知和语言的人身物质基础的研究。这一认知语言学研究分支的理论基础为“概念的体现”的思想。与此相应，人类概念世界（ 及自然语言的语义）的构成，至少包括某些最抽象的片断，受制于人的生物特性和身体与社会相互作用而取得的经验。

认知语言学研究的主要内容

7、对语言中隐喻和转喻/换喻的研究。隐喻的研究是莱考夫教授的“拿手好菜”，正是他将隐喻从一个传统的问题变成了语言学研究领域及诸多相关科学中一个相当时髦的一个话题。1980年，美国芝加哥大学出版的莱考夫和约翰逊的《我们赖以生存的隐喻》从隐喻的角度探讨语言的本质，用大量的语言事实证明语言与隐喻认知结构的密切相关性，将隐喻看作用不同领域的术语思考新的概念领域的工具。在莱考夫等志同道合者看来，隐喻决不仅仅是一种普通的语言现象，从根本上讲，隐喻是一种认知现象。隐喻性思维是人类认识事物、建立概念系统的一条必由之路。

内容提要

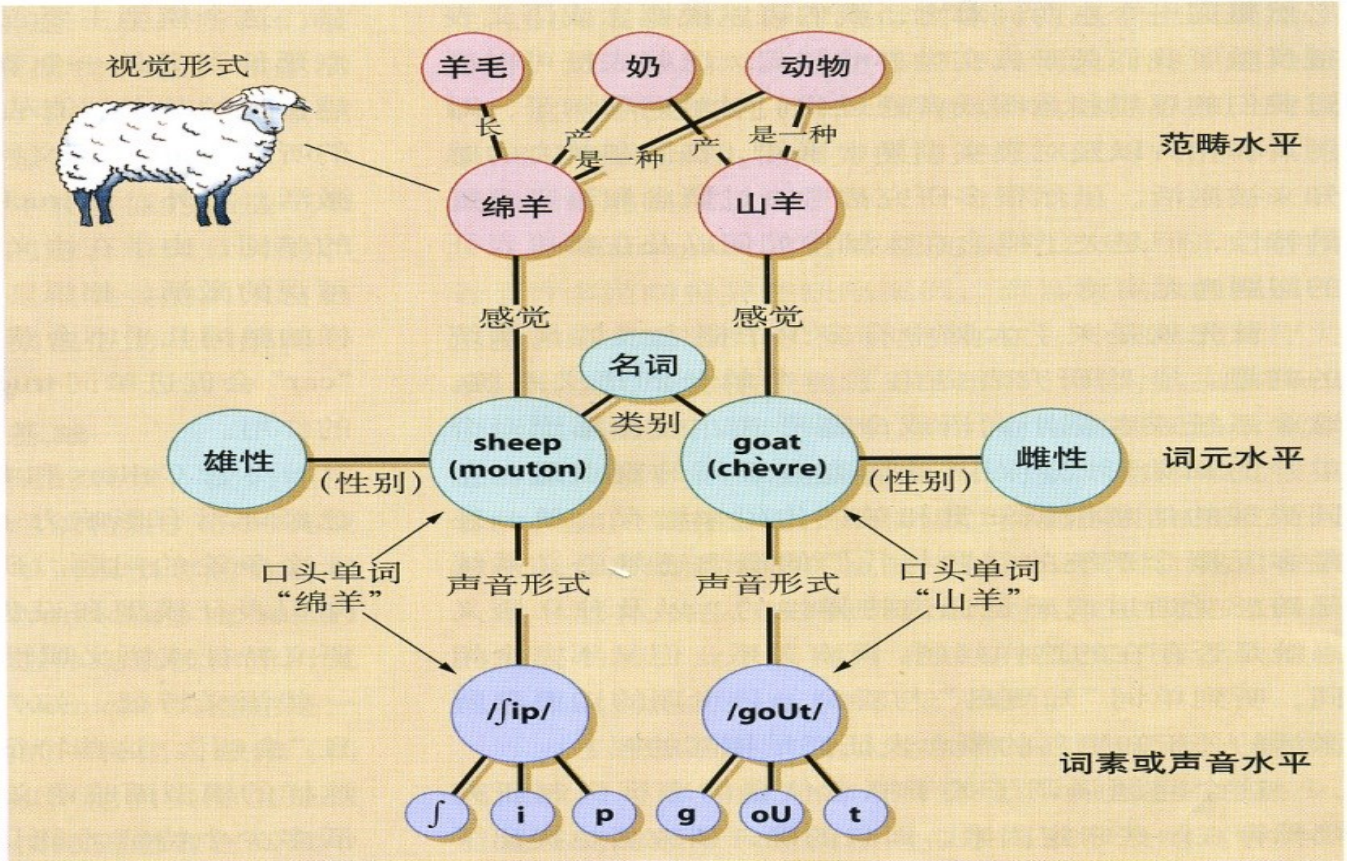
- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

心理词典

心理词典和一般的大学词典不一样，心理词典是以特异性信息网络的形式组织起来。荷兰心理语言学家列弗特(Levelt W)等提出特异性信息网络在所谓的词素(lexeme)水平上以单词的形式存在，在词元(lemma)水平上以单词的语法特性的形式存在。在词元水平上，单词的语义特性也被表征出来了。这种语义信息定义了概念水平。在这种概念水平下，使用某一特定的单词是适当的。

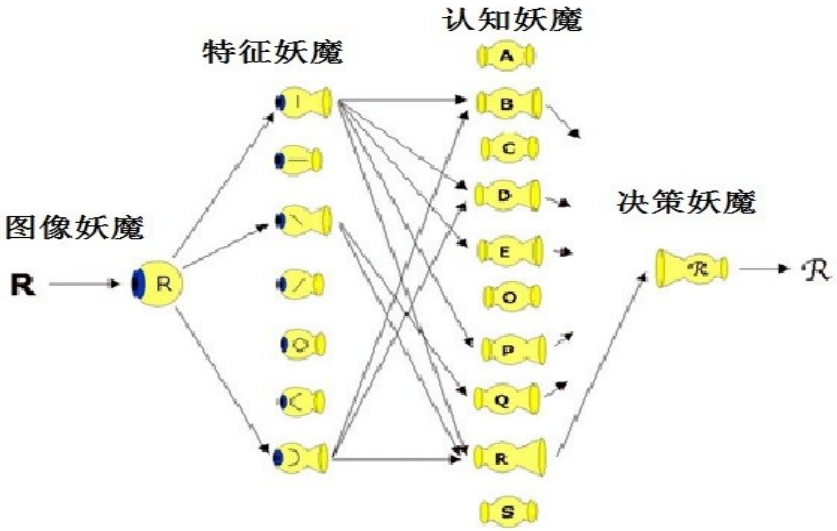
心理词典

词汇网路片段的例子



妖魔模型

1959年，塞尔弗里奇(Selfridge O G) 提出“妖魔模型”(Pandemonium model)。这个模型以特征分析为基础，将模式识别过程分为4个层次，每个层次都有一些“妖魔”来执行某个特定的任务，这些层次顺序地进行工作，最后达到对模式的识别。



心智的模块性

1983年，福多(Fodor J A)出版了《心智的模块性》，正式提出了模块理论。福多认为模块化结构的输入系统应该有以下特征：

（1）领域特异性。输入系统接收来自不同感觉系统的信息，用特异于系统的编码加工这些信息。例如，语言输入系统将视觉输入转化成语音或口语声音表征。

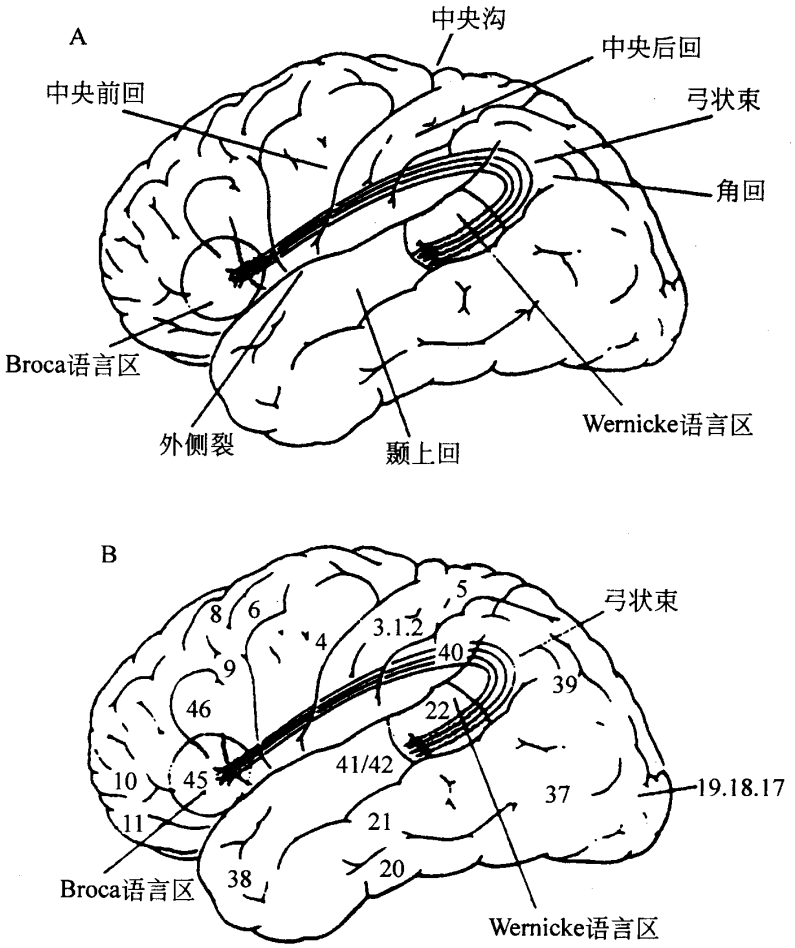
（2）信息封装。加工是严格朝一个方向进行的，不完整信息是不能够被传递的。在语言加工中不存在自上而下的影响。

（3）功能定位。每个模块是在一个特定脑区中实施功能的。

语言理解的神经模型

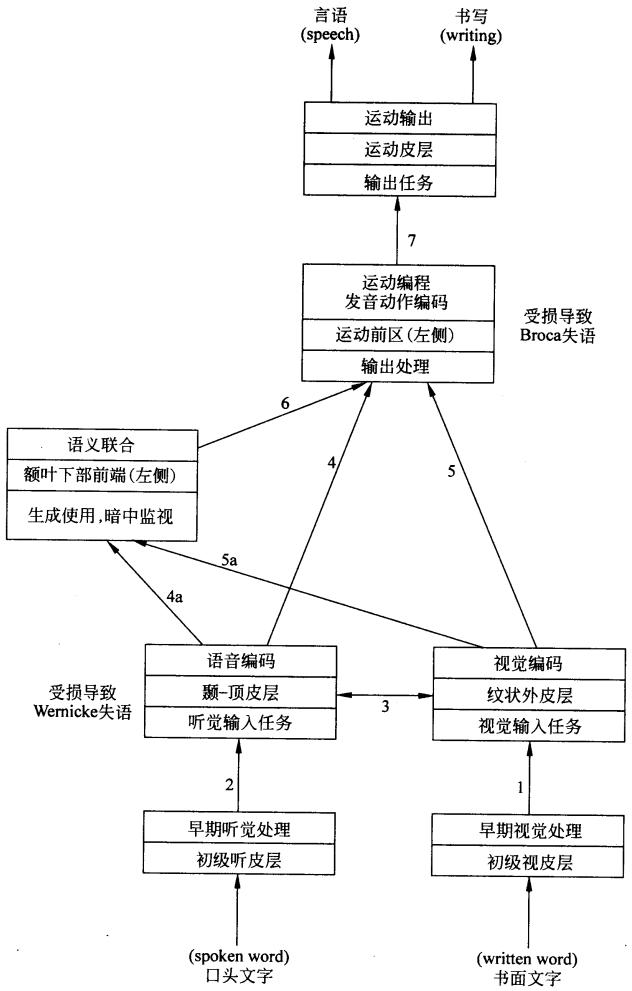
- 布洛卡（Broca P）通过失语病人的研究，提出产生口语的区域在左半球额叶下回。这个区域后来被称为Broca区。即Broca区。该区也称为前说话区，常常描述为额下回后1/3。用于计划和执行说话，病变损伤该区会导致运动性失语，主要表现为口语表达障碍。
- 在19世纪70年代，韦尼克（Wernicke C）提出Wernicke区，该区损伤后说话流利，但说出的是无意义的声音、词和句子，而且他们在理解话语时有严重困难。韦尼克发现听觉加工发生在颞横回内的颞上回前部，这个区域参与单词的听觉记忆，也就是单词的听觉记忆区。

大脑的语言区分布


A
中央沟
中央后回
中央前回
弓状束
角回
Broca语言区
外侧裂
颞上回
Wernicke语言区
B
弓状束
8
6
5
3.1.2
4
40
9
39
46
22
41/42
10
45
37
19.18.17
11
21
20
38
Broca语言区
Wernicke语言区

语言理解的神经模型

听觉输入在听觉系统被转换，然后信息传递到以角回为中心的顶颞枕联合皮质，再传递到Wernicke区，并在这里可以从语音信息中提取出单词表征。信息流从Wernicke区经过弓形束(白质神经束)到达Broca区，这里是语法特征记忆之所，同时短语结构可以在这里得到分配。接着单词表征激活概念中心相关的概念。这样，听觉理解就发生了。在口语产生中，除了概念区激活的概念在Wernicke区产生单词的语音表征，并被传递到Broca区来组织口语发音动作外，其他过程都是类似的。



内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

语法分析

- 语法分析的主要任务：
  - 确定输入句子的结构：识别句子的各个成分及其之间的关系
  - 句子结构的规范化：目的是简化后续处理
- 分析自然语言的方法主要分为两类：

  基于规则的方法：如短语结构语法和Chomsky语法体系

  基于统计的方法

短语结构语言

- 定义
  - 句子：一个符号串
  - 语言：句子的集合
  - 语法：对一个句集一种有限的形式化描述
- 描述一般语言的方法：
  - 识别器：由程序判断读入的符号串是不是一个句子
  - 短语结构语法：一种基于产生式的形式化工具，也称为产生式语法

短语结构语言

定义：短语结构语法定义为：G＝(T，N，S，P)

- T是终结符集合，即被定义的语言的所有词 汇（或符号）
- N是非终结符集合，这些符号用于描述语法 成分，并不出现于句子中。

  则有：V＝T∪N，T∩N＝Φ（空集），V是属于该语法的全部符号。
- S是起始符号，它是N中的一个成员。
- P是一个产生式规则集。$a \rightarrow b$　　$(a \neq b, a \in V^{+}, b \in V*)$

短语结构语言

- 在短语结构语法中，基本运算是把一个符号串重写为另一个符号串，每条语法规则也叫重写规则
- 一个句子的产生就是从S符号到词汇串的推导过程
- 如果一个程序能够根据一个短语结构语法来确定一个句子的推导，则它可称为一个句法分析器(parser)。
- 语法G所定义的语言记为L(G)：

$$L(G)=\{W \mid W\in T^*,\ S\Rightarrow^*_G W\}$$

短语结构语言

- 正则语法：
  - 正则语法有两种形式：
    - 左线性语法：如 $A \rightarrow a \mid Ba$
    - 右线性语法： 如 $A \rightarrow a \mid aB$
  - 可以表示如下的句子：
    - a*b*
  - 语法例子：
    - $S \rightarrow a \mid S_1 \mid a\ S$
    - $S_1 \rightarrow b \mid b\ S_1$
  - 与有限状态机等价

短语结构语言

- 上下文无关语法：
  - 语法规则形式为： $A \rightarrow \mathrm{x}$

    即左边为一非终结符，右边没有限制
  - 可以表示的句子如：
    - $\mathrm{a}^n\mathrm{b}^n$
  - 语法例子：
    - $S \rightarrow \mathrm{a} \mid S\ \mathrm{b}\ S$
  - 该文法应用于程序设计语言中

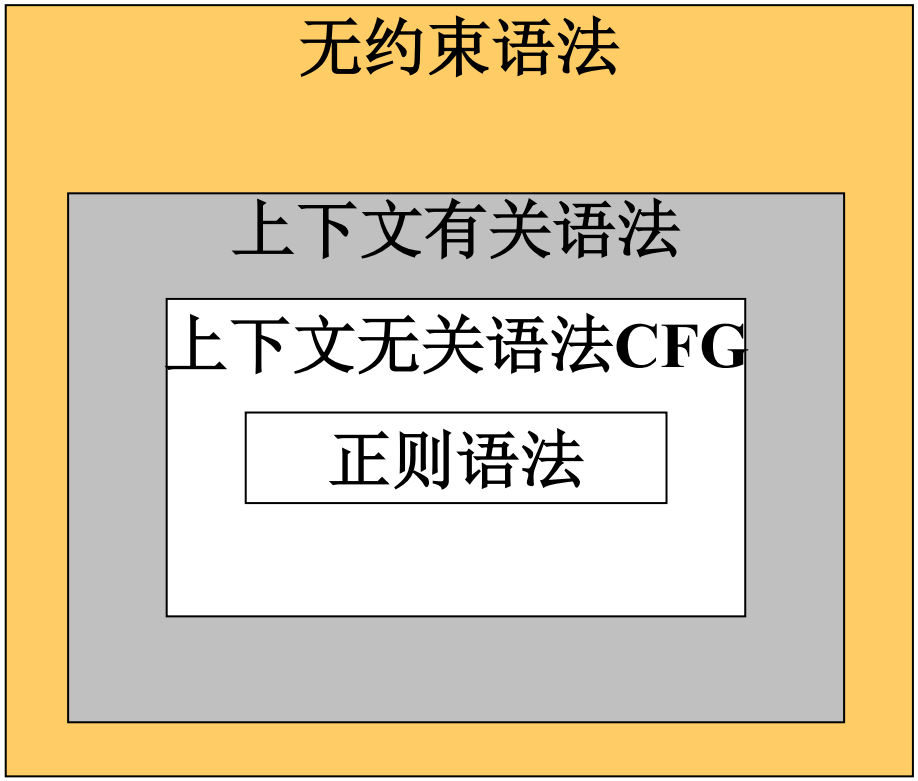
短语结构语言

- 上下文有关语法：
  - 语法规则：
    - 规则右边的符号数不能少于左边符号数
    - 右边的符号可以是终止符也可以是非终止符
  - 上下文有关语言是递归的
  - 可以表示的语言：
    - $a^n b^n c^n$
  - 语法例子：
    - $AB \rightarrow BA$

短语结构语言

- 无约束短语结构语法：
  - 语法规则是没有限制的：
    - 左边可以是任意多个终止符或非终止符
    - 右边可以是任意多个终止符或非终止符
  - 该语言是递归可枚举的
  - 该语言与图灵机等价
  - 语法例子：
    - $A\ B \rightarrow C$

Chomsky形式文法体系



乔姆斯基

CFG的分析算法

- 用一个短语结构语法对一个句子进行语法分析，意味着寻找一个从起始符到该句子的推导，这个推导一般可以表示为一棵句法树
- 一般一棵句法树对应的推导不是唯一的，但是如果在推导过程中每次总是重写最左边的非终止符，则称该推导为最左推导。
- <SENTENCE> ⇒<SUBJECT><VERBPHRASE>

⇒Mary <VERBPHRASE>

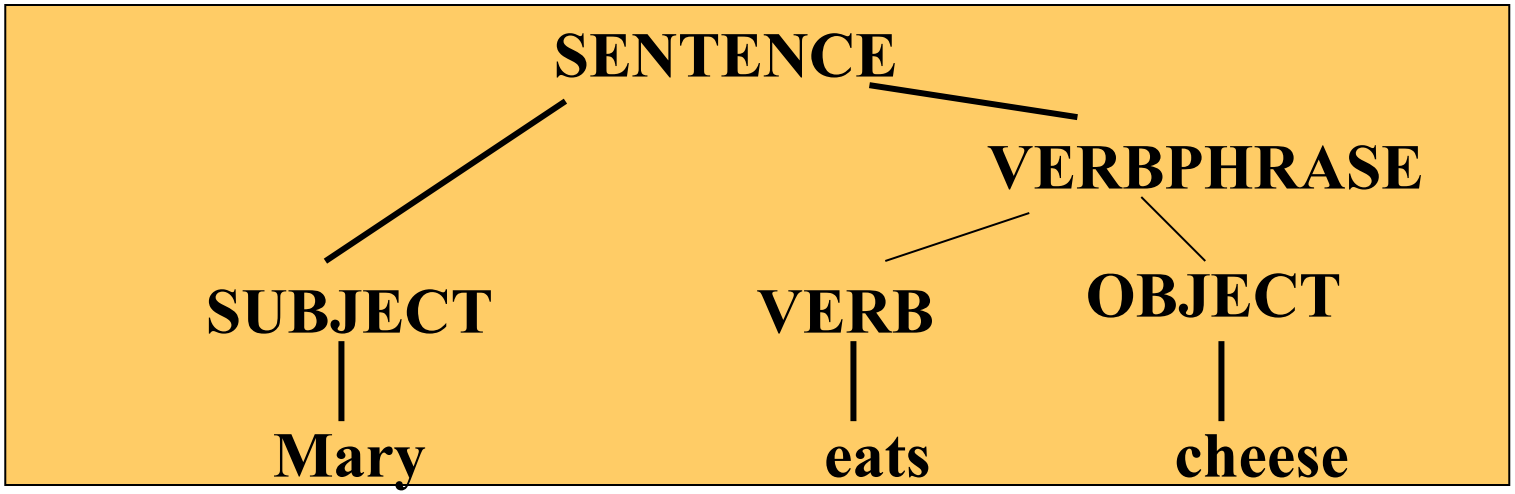
⇒Mary <VERB> <OBJECT>

⇒Mary eats <OBJECT>

⇒Mary eats cheese

CFG的分析算法

<SENTENCE> ::=<SUBJECT><VERBPHRASE>
<SUBJECT> ::=John | Marry
< VERBPHRASE>::=<VERB><OBJECT>
<VERB> ::=eats | drinks
<OBJECT> ::=wine | cheese



CFG的分析算法

- 句法分析器分为：
  - 从推导方向来分：
    - 自顶向下：从树顶的根结点开始推导建立句法树，方向是从起始符S到句子
    - 自底向上：从树底部的叶结点(词或词类)规约，建立句法树，方向是从句子到S
  - 从算法上分：
    - 回溯算法：每次只尝试一种推导，当这种推导失败时便返回以尝试另一种推导
    - 并行算法：同时进行所有的推导

CFG的分析算法

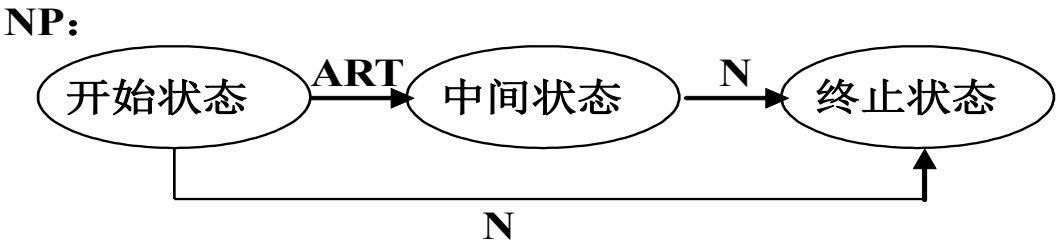
- 自顶向下的回溯算法
  - 该方法逐个地枚举推导直到找到一个能生成句子的推导
  - 一般，对具有左递归的语法，该方法需要增加某些测试以避免陷入死循环
  - 对于”Mary eats cheese”的句法和推导为：

S→NP+VP	(1)	S⇒NP+VP	(1)
NP →N	(2)	⇒N+VP	(2)
VP →V	(3)	⇒N+V+NP	(4)
VP →V+NP	(4)	⇒N+V+N	(2)

转移网络

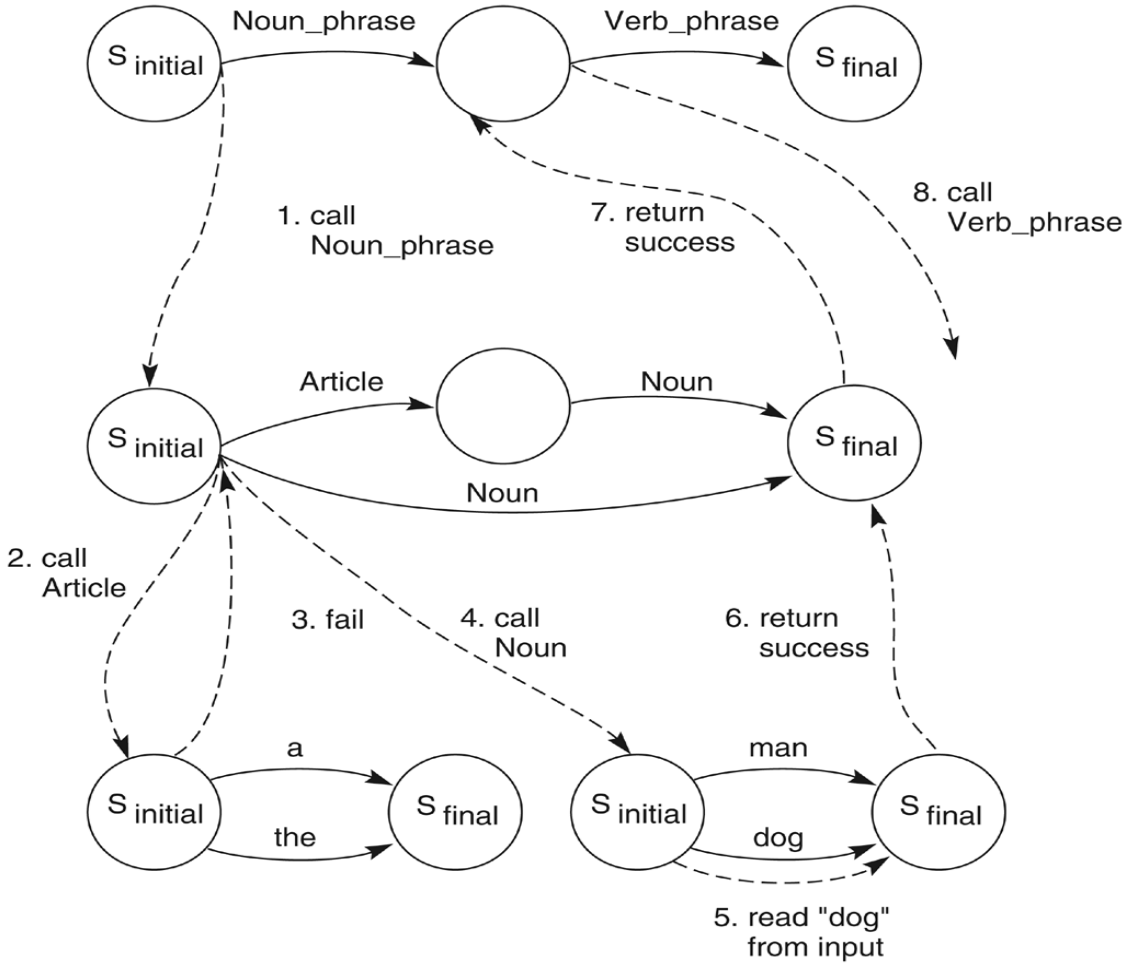
- 转移网络在自动机理论中用来表示语法。
- 句法分析中的转移网络由结点和带有标记的弧组成，结点表示状态，弧对应于符号，基于该符号，可以实现从一个给定的状态转移到另一个状态。

图 16.2(a) S→NP+VP 的转移网络



NP→ART+N 和 NP→N 的转移网络

Dog bites


S initial
Noun_phrase
Verb_phrase
S final
1. call
Noun_phrase
7. return
success
8. call
Verb_phrase
S initial
Article
Noun
S final
Noun
2. call
Article
3. fail
4. call
Noun
6. return
success
S initial
a
the
S final
S initial
man
dog
S final
5. read "dog"
from input

扩充转移网络

扩充转移网络ATN

- ATN是20世纪70年代由W. Woods提出来的
- ATN语法属于一种增强型的上下文无关语法，即用上下文无关文法描述句子文法结构，并同时提供有效的方式将各种理解语句所需要的知识加到分析系统中，以增强分析功能，从而使得应用ATN的句法分析程序具有分析上下文有关语言的能力。
- ATN主要是对转移网络中的弧附加了过程而得到的。当通过一个弧的时候，附加在该弧上的过程就会被执行。这些过程的主要功能是（I）对文法特征进行赋值；(II)检查数（number）或人称（第一、二或三人称）条件是否满足，并据此允许或不允许转移。

内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

认知语义学

- 认知语义学强调人类在认知过程中与周围世界的相互作用，认为基本类概念和图像一图式是人类建立复杂认知模式的根本。概念的建立与人自身的经验关系密切。
- 认知语义学认为，概念经验的二类结构是人类认识和理解周围环境和抽象概念的基础和必要条件：
  - 基本类范畴，在感知上具有总体的形状和单一的心理影像，并具有容易和快速辨认的特征
  - 图像一图式是人类理解的另一基础，例如容器、力、路径、连接、上下、前后、部分一整体等。

语义分析

- 语义分析的任务：

  输入句子的句法结构和句子中每个实词的词义推导出能反映该句子意义的某种形式化表示

- 对语义现象作形式化处理要比句法现象困难得多，主要原因有

  -语义和句法系统的界限很难划清楚

  -语义及其他认知系统的界限也难以划清楚。

  -用于计算机语义处理的计算语义学还远未成熟

格文法

- 格语法是Filmore于1968年提出来的，曾 经对自然语言理解技术的发展产生过较 大的影响，直到现在不少研究仍在使用格语法。因为人们认识到格关系确实是描述语言语义(包括和语法的关系)的一种很好的形式，当然在实际应用过程中不可避免地要有些修改。

格文法

- 格文法的特点是允许以动词为中心构造分析结果，尽管文法规则只描述句法，但分析结果产生的结构却相应于语义关系，而非严格的句法关系
- 如句子：Mary hit Bill

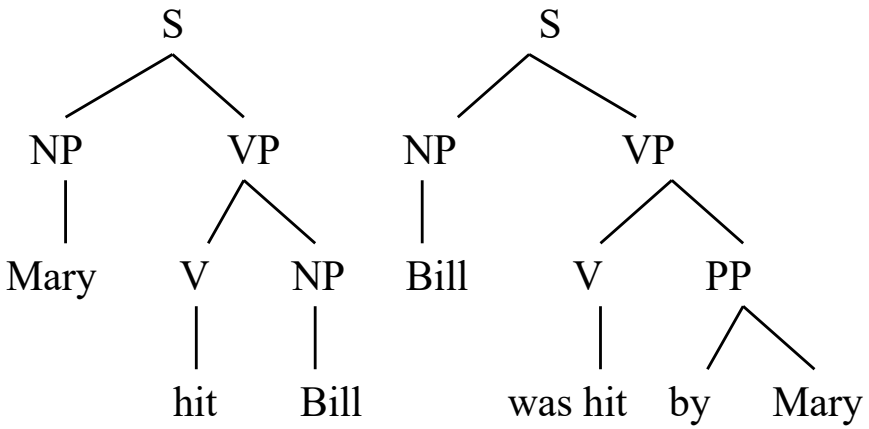
  的格文法分析结果可以表示为

  (hit (Agent Mary)

  (Dative Bill))
- 在格文法中，格表示的语义方面的关系，反映的是句子中包含的思想、观念等，称为深层格。和短语结构语法相比，格文法对于句子的深层语义有着更好的描述。

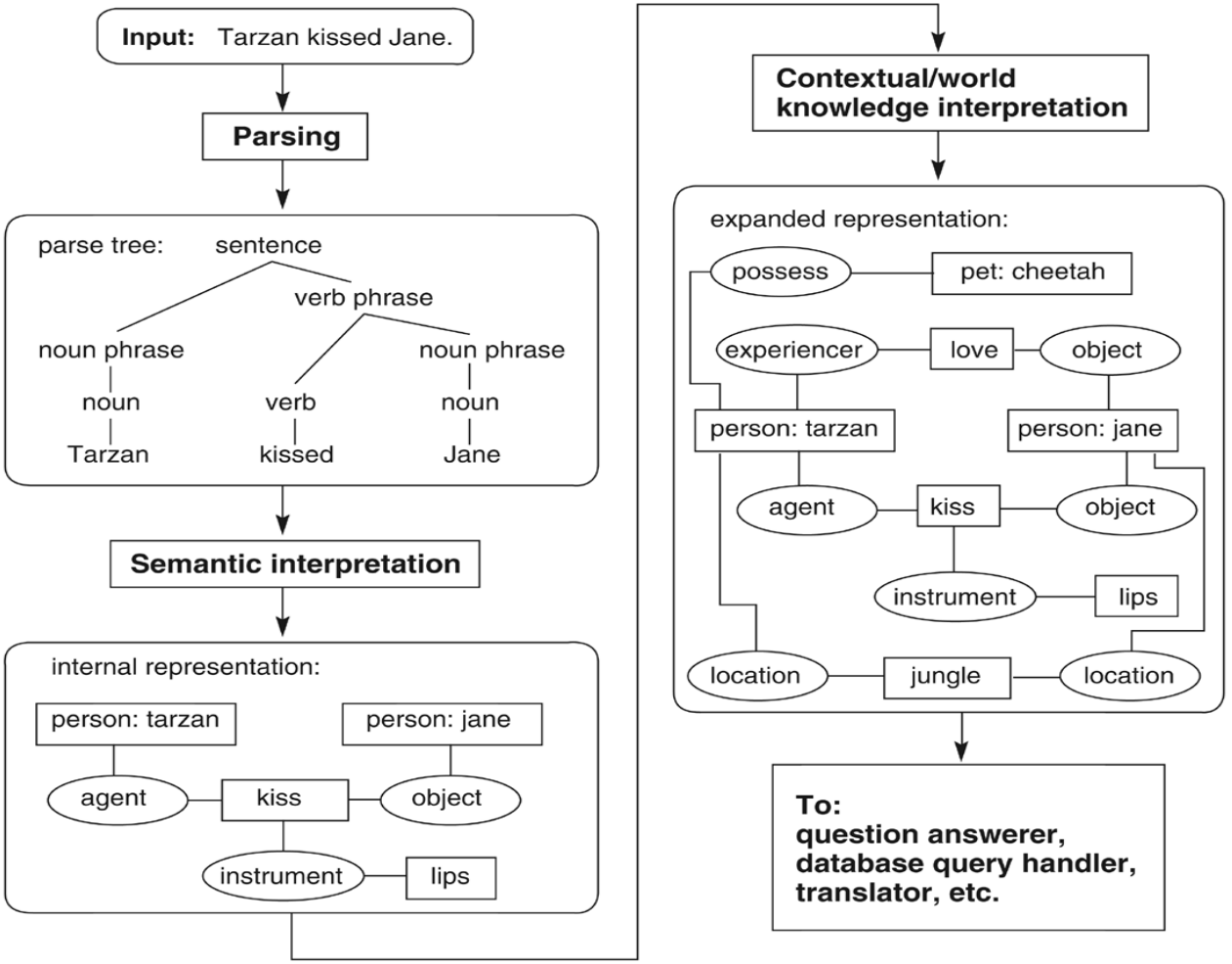
格文法

- 如果两个句子的底层的语义关系一致，各名词成分所代表的格关系不会发生相应的变化。例如，被动句“Bill was hit by Mary”与上述主动句具有不同的句法分析树，但格表示完全相同。



主动句和被动句的句法分析树

格文法


Input: Tarzan kissed Jane.
Parsing
parse tree: sentence
verb phrase
noun phrase
noun phrase
noun
verb
noun
Tarzan
kissed
Jane
Semantic interpretation
internal representation:
person: tarzan
person: jane
agent
kiss
object
instrument
lips
Contextual/world knowledge interpretation
expanded representation:
possess
pet: cheetah
experiencer
love
object
person: tarzan
person: jane
agent
kiss
object
instrument
lips
location
jungle
location
To:
question answerer,
database query handler,
translator, etc.

语用分析

- 语用分析与知识、上下文和推理等因素有关。维诺格拉德（Winograd T）认为语言是一个讲话者和听者之间关于一个共同的世界的一种通信手段。语言是一种社会交际工具，研究语言必须研究其社会功能。维诺格拉德认为语义理论必须在三个平面上描述关系，

  （1）确定词的意义

  （2）确定词组在句法结构中的意义

  （3）一个自然语言的句子决不应被孤立地解释。

- 一种语义理论必须描述一个句子的意义如何依赖于它的上下文。

语用分析

- 语义理论必须涉及语言学背景（说话的上下文）和现实社会背景（即同非语言学事实的知识的相互作用），语义理论必须同句法和语言的逻辑方面（演绎推理）相联系。正是基于这些观点，即语法、语义和语用学相互作用的观点，1970年维诺格拉德成功地研究了被人称为“绝技”的自然语言对话系统SHRDLU，实现人与计算机之间的灵活对话。这项创举震动了当时的人工智能界。

语言象似性

- 语言象似性（iconicity）被定义为语言符号在语音、语形、结构上与其所指之间存在映照性相似的现象，即讨论形式与意义之间的关系。象似性研究为当代语言学开辟了一个崭新的领域，能够更加深刻的认识到形成语言规则背后的认知机制。
- 语言的象似性有距离象似性、数量象似性、顺序象似性、标记象似性等。
- 象似性理论与语用原则存在共同之处，可以将象似性理论于语用学结合起来研究。

内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

隐喻与转喻

- 认知语言学家认为，语言能力是一般认知能力的反映，并由一般的神经过程所控制。根据这一观点，各种认知之间是一个连续体，而语言不是人的心灵和大脑中独立的“模块”。认知语言学家从神经学和认知心理学中证明了这一观点。
- 在各种认知能力中，一个主要的和普遍的认知能力是想象(imagination)，即把一些概念投射到另一些概念中去。这就是为什么想象机制的隐喻和转喻会成为认知科学家研究的重点之一。

隐喻与转喻

- 隐喻是一个认知机制，在这一机制中，一个认知域被部分地映现(mapped)于另一认知域上，后者由前者而得到部分地理解。前者叫来源域(source domain)，后者叫目标域(target domain)。
- 转喻是在一个认知域中映现，如部分代表整体就是一例 。
- 隐喻和转喻一些最抽象和重要的隐喻和转喻可作为基本的来源域，如一些普遍的空间概念（垂直性和包容性等），它们被称为图象图式。这些图象图式是基于人的最基本的身体经验而习得。

隐喻与转喻

- 隐喻和转喻都是认知模式的基本类型，两者都以经验为理据，并用于某些语用目的。把隐喻和转喻作为“模式”强调了它作为稳定的“认知装备”（cognitive equipment）的一部分，即隐喻和转喻应是我们人类范畴系统的稳定成分。
- 近些年来，把隐喻和转喻看作概念整合的一个特例。概念整合理论与隐喻和转喻的双域理论并不矛盾，因为前者以后者为前提。然而概念整合理论能更准确地解释隐喻和转喻的运作情况，而且还能解释隐喻和转喻的认知理论解释不了的现象。

隐喻与转喻

- 隐喻和转喻常常相互作用，有时异常复杂，其相互作用的方式有两种类型 ：

（1）在纯粹概念层次上相互作用；

（2）在同一语言词语中，隐喻和转喻在话语中的相互示例。

- Barcelona认为在这两种类型中，第一种类型最重要，并有两种次类型：
  - ➢隐喻的转喻理据
  - ➢转喻的隐喻理据

转喻研究的问题

- 认知语言学早期重视隐喻研究，现在逐渐重视转喻研究，因为人们发现转喻与隐喻比起来在概念上是更基本的一种认知活动。
- 目前转喻的研究主要集中在以下三点：
  - 转喻是一种概念映现还是一个域中的概念激活?
  - 转喻总是指称性的吗？
  - 转喻是如何成为常规化语言的？

  例如转喻“John is a Picasso.”就不是指称的，而是喻指John是画画的天才。

内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

心理空间理论

- 1985年，G. Fauconnier在其著作《心理空间》中提出了心理空间理论，系统地考察人类认知结构和人类语言结构在认知结构中的体现。
- 心理空间理论是意义建构的理论，包含句子意义是如何被分割成心理空间。虽然该理论都是处理语言材料，但它在本质上不是语言的。心理空间是说话人谈论实体和其各种关系时建构的一些可能世界和有关某一领域的信息集合。

心理空间理论

- 心理空间理论认为，语言结构的基本功能是利用描写认知视角的不同的信息辨认度(accessibility）来考察语言的用法 。
- 心理空间的各种连接或映现可使我们使用词语作为触发词(trigger)去指称其它心理空间中的另一目标实体，这些连接或映现包括语用功能，转喻、隐喻和类比等。语用功能可把两个心理空间连接起来，例如作者名字可与该作者所著的书对应起来。

心理空间理论

- 心理空间建构和连接的基本思想是，当我们思维和谈话时，在语法、语境和文化的压力下，建构和连接心理空间。随着话语的展开，我们创造出一个心理空间网络。由于每个空间都来自于一个母空间(parent space)，而每个空间又有许多子空间，所以空间网络将是个二维点阵(two-dimensional lattice)。在这个空间网络中，我们可以从子空间到母空间，也可以从母空间到子空间。

心理空间理论

意义建构的动态过程包括三点：

- 在话语的某一点上，建立并连接心理空间，其中一个空间是用来表示视点(viewpoint)，即该空间是辨认其它空间的起点；
- 某一特定空间是话语的焦点(focus)；
- 在空间网络中移动是从基础空间(base space)开始，它提出了最初的视角，然后使用合适的空间连接词(connector)变换视角和焦点。

概念整合理论

随着心理空间理论的发展， Fauconnier 和Mark Turner先后发现了反映许多语言现象中的一条重要的心理空间的认知操作：概念整合(conceptual blending)。概念整合包括建立相互映现的心理空间网络，并以各种方式整合成新的空间。

概念整合理论

基本的概念整合网络包含四个心理空间。其中两个称为输入空间(input spaces)，并在其之间建立跨空间的映现。跨空间映现创造或反映了两个输入空间所共享的更抽象的空间，即类属空间(generic space)。第四个空间是整合空间(blended space)，是从输入空间中进行选择性的映现而来的，它可以各种方式形成两个输入空间所不具备的突生结构(emergent structure)，并可把这一结构映现回网络的其它空间中去。

内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

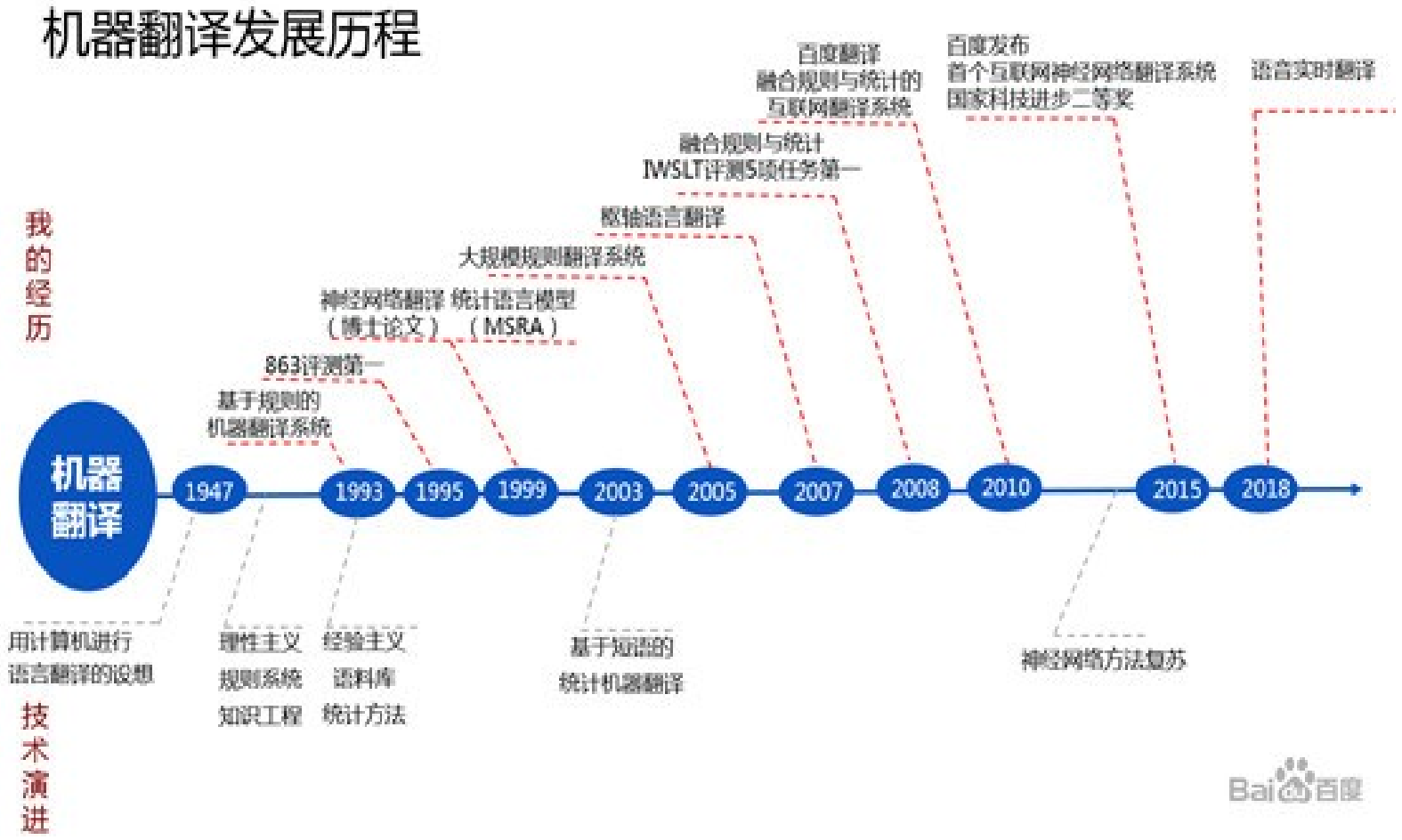
机器翻译

- 机器翻译是利用计算机把一种自然语言转变成另一种自然语言的过程。用以完成这一过程的软件叫做机器翻译系统。机器翻译是语言学、人工智能、计算技术、认知科学等学科相结合的产物。
- 人作翻译时，把一个源语言句子译成目标语言句子，涉及到四个基本操作：目标语言单词的检索、调序、删词、增词；机器翻译系统的操作过程也不例外，有检索、分析、转换和生成的主要四个阶段。这被称为基于分析和转换的机器翻译系统。也被认为是模拟人类翻译活动最恰当的机制。

机器翻译

- 20世纪50年代初到60年代中为大发展时期。但是由于当时对机器翻译的复杂性认识不足而产生了过分的乐观情
- 20世纪60年代中到70年代初由于遇到了困难而处于低潮时期。
- 20世纪80年代机器翻译开始复兴，注意力几乎都集中在人助自动翻译上，人助工作包括译前编辑（或受限语言），翻译期间的交互式解决问题，译后编辑等。几乎所有的研究活动都致力于在传统的基于规则和“中间语言”模式的基础上进行语言分析和生成方法的探索，这些方法都伴有人工智能类型的知识库。
- 在20世纪90年代早期，机器翻译研究被新兴的基于语料库的方法向前推进，出现新的统计方法的引入以及基于案例的机器翻译等。

机器翻译

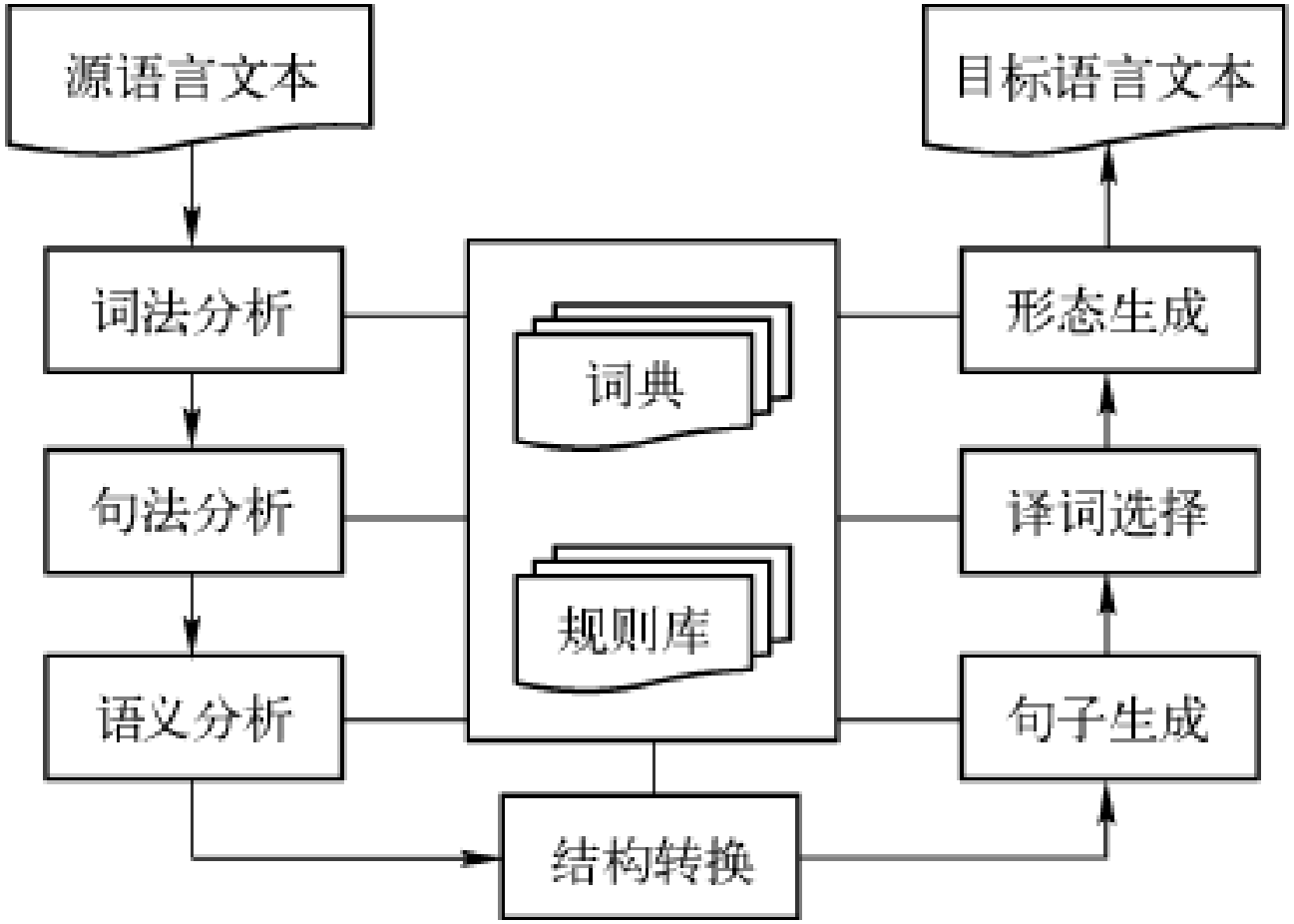


引自百度王海峰的报告

机器翻译

- 机器翻译的一般过程包括：源语文输入、识别与分析、生成与综合和目标语言输出。当源语文通过键盘或扫描器或话筒输入计算机后，计算机首先对一个单词逐一识别，再按照标点符号和一些特征词(往往是虚词)识别句法和语义。然后查找机器内存储的词典和句法表、语义表，把这些加工后的语文信息传输到规则系统中去。从源语文输入的字符系列的表层结构分析到深层结构，在机器内部就得到一种类似乔姆斯基语法分析的“树形图”。

基于规则的机器翻译流程图



基于统计的机器翻译方法

基于统计的机器翻译方法

基于统计的机器翻译方法，一般不要任何语言学知识，它的基本原理是实现源语言词汇到目标语言词汇的映射。其思路受到语音识别研究的启发，因而应用了类似的方法来实现。研究者用大规模的双语语料库作为基础，对源语言和目标语言词汇的对应关系进行统计，根据统计规律输出译文。这种方法没有使用语言知识，主要特征是概率统计与随机过程的方法成为了分析和生成过程的唯一方法。它的主要内容是双语句对的对齐，通过词汇同现的可能性来计算一种语言的一个词映射到另一种语言的一个词（或两个、零个词）的概率。应该说，基于统计的机器翻译方法的出现改变了机器翻译研究的面貌，从而开始了机器翻译研究的新阶段。不过，有些学者也对纯统计方法提出了异议，认为必须引入高层语法、语义模型，显然这是正确的。否则，基于统计的机器翻译方法不可能产生高质量。

基于统计的机器翻译方法

由于当前计算机在运算速度和存储容量方面都有巨大的提高，可以获取大量的机读语料库，因此在机器翻译中应用统计方法的条件已经成熟。

机器翻译的噪声通道模型可以视作最早的机器翻译思想的某中复活，其思路可以这样理解：

假设说话者已经用目标语想好了一句话T，但是说出的却是源语言句子S。这样一个过程可以看成为编码过程。而统计的机器翻译就是要从S回推T，可以看成解码过程。这样，统计的机器翻译任务分为两个部分：一是建模，即建立翻译的计算方法以及从双语语料库中估计模型的参数；二是解码，即寻求一种高效搜索算法取有关概率计算的最大值。

基于统计的机器翻译方法

在概率计算的前后，转换是进行有关预处理和后处理，例如句子当中的日期、时间、数字、人名等应该用适当的类别标志加以替换，以便更好地进行计算，计算后再换回来。因为这类词不属于一般的词汇，单词本身在语料库中的出现缺乏代表性。

建模就是设计各个模型的计算公式。因为直接计算某个句子出现的概率是比较困难的，语料库不可能足够大到包含所有句子，必须进行合理的、适当的简化。这是统计方法的特点，所得到的结果是近似值，但是因为概率本身就不是精确的，所以这些近似完全可以接受。

基于统计的机器翻译方法

总之，基于统计的机器翻译方法可以简单的这样看：将原始的某个句子按词折开，然后全部单词存储；翻译则是取出，按概率统计的方法重组句子，这样的句子就是统计方法的翻译结果。

当然，我们还是认为应该加强统计方法与语法分析、语义分析相结合的研究。

基于记忆的机器翻译方法

- 建立机器翻译系统需要大量的知识。在基于转换和基于中间语言的机器翻译方法中，知识按一定规则译成代码，这既耗时花费也大。此外，知识获取瓶颈阻碍了机器翻译的快速发展，这是早期机器翻译面临的重大难题。
- 为克服这一困难，除了上节介绍的统计方法外，日本机器翻译专家Nagao在80年代提出了一种新方法，用已经存在的翻译实例（双语文本）作为知识源，这种方法称为基于记忆的翻译，后来普遍称为基于实例的翻译。基于实例的思想已被广泛的采用，既用于机器翻译的设计，也用于机器翻译不同处理阶段的实现中。用不断积累的已经译好的文本作为机器翻译的样本的思想，也是具有吸引力的。

基于记忆的机器翻译方法

基于记忆的机器翻译方法通过结构化的翻译例子直接把源语言的短语和句子与目标语言的短语和句子对应起来。方法的不同使得处理步骤或多或少，但都必须实现源语言到目标语言的转换，其映射关系或者是词到词，或者是短语或句子到与之相应的等价物，或者是一棵句法树到另一棵句法树。

基于记忆的机器翻译（EBMT）的实现过程简单概述如下：给定源语言输入句子S，在双语语料库C中匹配查找一个最相近的句子S’， 则S’的译文T就被接受为S的译文。

基于记忆的机器翻译方法

翻译的过程一般就是查找和复现相似的例子，发现和记起特定的源语言表达或相似的表达在以前是如何翻译的，把以前的翻译实例作为主要知识源。

基于记忆的机器翻译方法的基本思想：

（1）把翻译实例存入翻译数据库。例如，存入汉语和英语句子对；

（2）对输入的句子，在翻译数据库中检索类似的翻译例句；

（3）调整实例后生成译文。

基于记忆的机器翻译方法

基于记忆的机器翻译方法可以按如下步骤实现：

（1）对双语语料库进行句子级对齐；

（2）在语料库的源语言一边进行句子分块，称为组块。然后检索输入组块的最佳匹配侯选，称为源语言内部匹配；

（3）在源语言最佳匹配后选的组块中检索对应目标语言组块，称为双语匹配；

（4）对组块级检索结果进行组合，以获得整个源语言文本的翻译结果。

基于记忆的机器翻译方法

源语言的内部匹配就是在语料库中查找一个与待译句子最相似的句子。对于任意给定的一个句子，很难在语料库中找到与之完全匹配的句子，所以对输入文本在语料库源语言一边的匹配查找采用了松弛匹配技术。松弛匹配就是部分匹配，不同的部分匹配被赋予不同的分值，以反映输入串和语料库中某些句子串的接近程度。整个输入组块与语料库组块匹配的分值由一定的公式计算出来。最后，待翻译文本中所有被匹配的每个输入组块都在语料库中检索出若干个最相近的组块，组块可以是一个句子、或者是一个从句、或者是一个短语，按照匹配分数从大到小排列。

基于记忆的机器翻译方法

源语言内部匹配的输出成为输入，即把前面从语料库中查到的组块、所在的句子和对应译文等构成当前处理步骤的输入。因为每个输入组块可能在语料库中匹配若干组块，所以每个组块有待进一步处理，即在双语匹配处理过程中又分为若干子过程，包括通过查词典获取词组译文；建立词汇级双语对照表；通过评分机制求出最佳翻译句等等。

基于记忆的机器翻译方法的输出结果带有相应的分数，这些分值来自不同的标准，相差可能很大。因此必须采用合适的函数（方式）对所有分数进行统一（归一化），然后这些分值才有可比性。

基于记忆的机器翻译方法

基于记忆的机器翻译方法的其它优点：

（1）可以通过索引和并行处理提高处理速度；

（2）可以采用最佳匹配推理；

（3）可以较好地利用翻译专家的专业知识（通过翻译实例）；

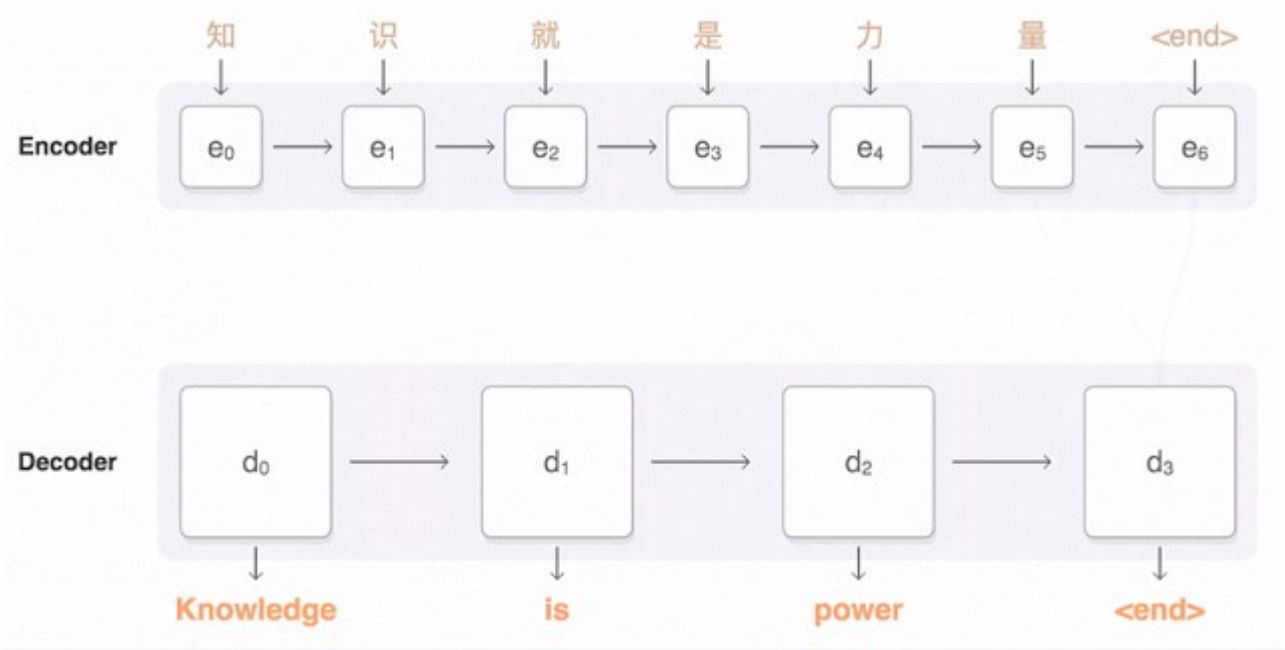
（4）一个基于实例的机器翻译系统的知识可以移植、共享。

基于深度学习的机器翻译方法

- 2016年9月28日，Google宣布发布Google神经网络机器翻译（GNMT：Google Neural Machine Translation）系统，该系统使用了当前最先进的训练技术，能够实现到目前为止机器翻译质量的最大提升。
- 将一个中文句子翻译成英语句子的过程。首先，该网络将这句中文的词编码成一个向量列表，其中每个向量都表示了到目前为止所有被读取到的词的含义（编码器“Encoder”）。一旦读取完整个句子，解码器就开始工作一一一次生成英语句子的一个词（解码器“Decoder”）。为了在每一步都生成翻译正确的词，解码器重点注意了与生成英语词最相关编码的中文向量的权重分布（注意“Attention”；蓝色连线的透明度表示解码器对一个被编码的词的注意程度）。

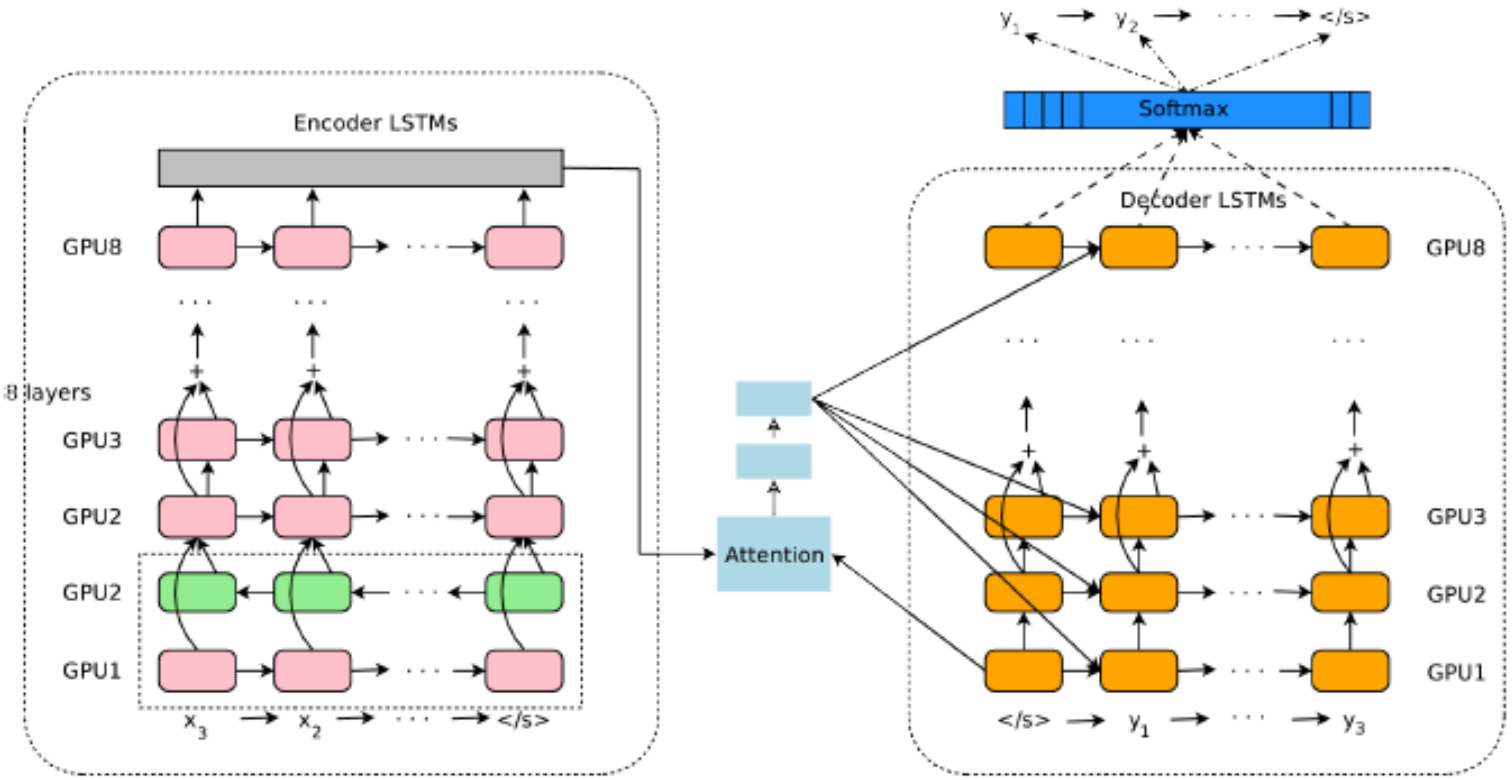
Google GNMT

Google GNMT



Google GNMT

Google GNMT



Google GNMT

- 移动版和网页版的 Google Translate 的中英翻译已经开始完全使用神经网络机器翻译系统了——每天大约 1800 万条翻译。其中，Google开放的机器学习工具套件 TensorFlow 和张量处理单元为部署强大的神经网络机器翻译系统模型提供了足够的计算力，同时也满足了 Google Translate 严格的延迟要求。
- Google神经网络机器翻译系统仍然会犯一些人类译者永远不会出的重大错误，例如漏词和错误翻译专有名词或罕见术语，以及将句子单独进行翻译而不考虑其段落或上下文。

内容提要

- 概述
- 语言处理认知模型
- 语法分析
- 认知语义学
- 隐喻和转喻
- 心理空间理论
- 机器翻译
- 问答系统

问答系统

- 问答系统(question answering system, QA)是信息检索系统的一种高级形式，它能用准确、简洁的自然语言回答用户用自然语言提出的问题。问答系统是目前人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向。2011年2月14日，在美国最受欢迎的智力问答节目《危险边缘》(Jeopardy)中，IBM的“沃森（Watson）”超级计算机击败该节目的两名总冠军詹宁斯（Jennings K）和鲁特尔（Rutter B），实现有史以来首次人机智力问答对决，并赢取高达100万美元的奖金。这是人工智能技术取得成功的代表。

危机边缘

2011年，“沃森”在电视娱乐节目《危机边缘》中战胜人类选手肯•詹宁斯和布拉德•拉特

问答系统

一般问答系统模型分为三层结构，分别为：用户层、中间层、数据层。各部分的主要功能如下：

（1）用户层（UI）：供用户输入提问的问题，并显示系统返回的答案。

（2）中间层（MI）：中间处理层，主要负责：分词、处理停用词、计算词语相似度、计算句子相似度，返回答案集。

（3）数据层（DI）：系统的知识库存储，主要有：专业词库、常用词库、同义词库、停用词库、课程领域本体、《知网》本体、常见问题集（FAQ） 库。

问答系统

- 问答系统自动答题的步骤如下：

1）根据专业词库，常用词库，同义词库对于用户输入的自然语言问句通过逆向最大匹配的方法进行分词，对于未登记词借助于分词工具把未登记词添加到词库中，在分词过程中同时标注词的词性和权值；

2）对于分词后的结果依据停用词库，并参考词性，删除停用词；

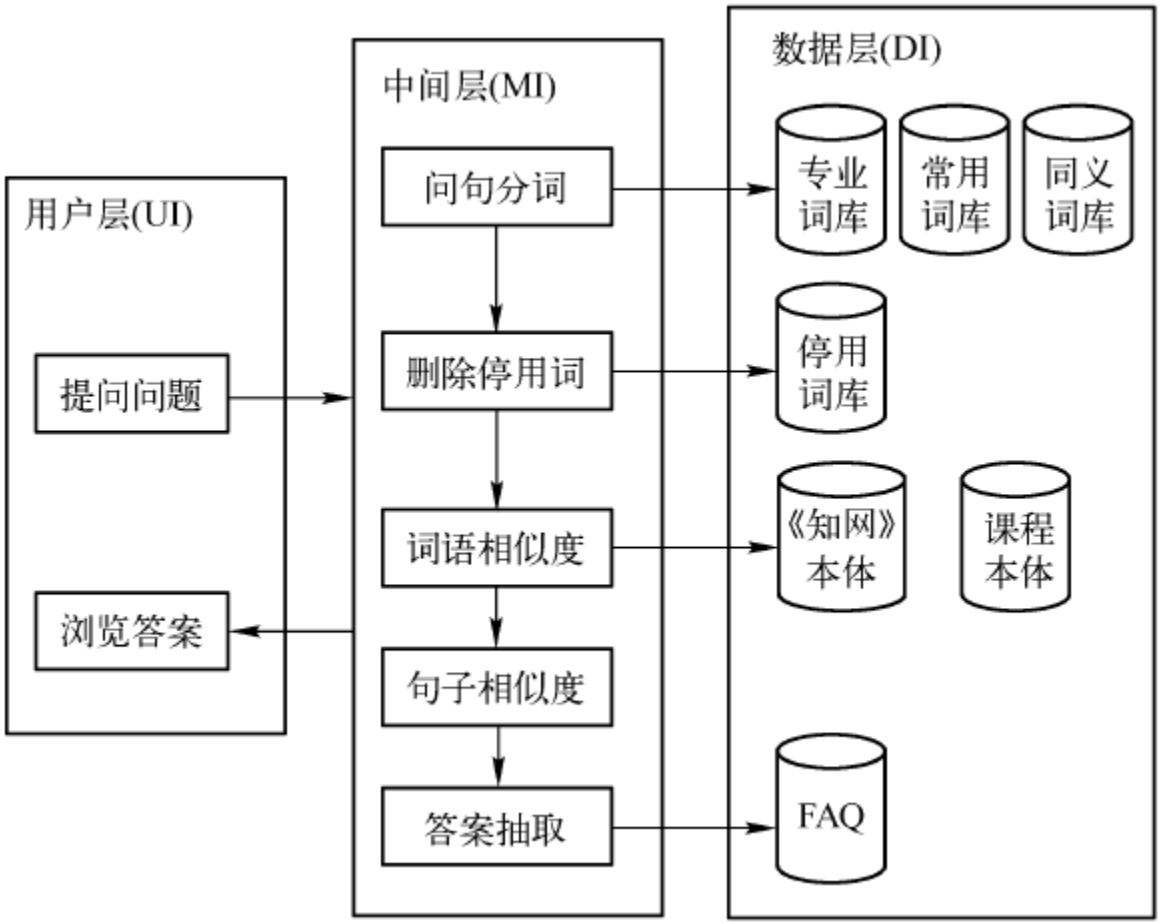
3）对于专业词汇采取基于本体的概念相似度方法进行计算词语语义相似度，对于其他词汇采取基于《知网》本体计算词语语义相似度；

问答系统

4） 分别计算IFIDF 相似度，根据词语的语义相似度来计算句子的语义相似度，计算词形、句长、词序、距离相似度来计算句子的结构相似度，最后组合起来加权求和计算句子相似度；（注：基于关键词向量空间模型的TFIDF 问句相似度计算方法是一种基于语料库中出现的关键词词频的统计方法，它是建立在大规模真实问句语料基础之上的。）

5）根据计算用户提问的问题与FAQ 中问题的句子相似度，定义一个相似度阈值，从FAQ 中抽取不小于相似度阈值且相似度最高的问题及其答案作为用户提问问题的答案；对于从FAQ 中抽取不到答案的问题通过发邮件给专家，添加到待解决问题集中，专家回答更新FAQ。

问答系统的结构框图


用户层(UI)
提问问题
浏览答案
中间层(MI)
问句分词
删除停用词
词语相似度
句子相似度
答案抽取
数据层(DI)
专业词库
常用词库
同义词库
停用词库
《知网》本体
课程本体
FAQ

检索算法

用户打开网页后在文本框中输入关键字进行搜索，系统将根据用户输入的关键字进行搜索，并返回和关键字相关的信息，若用户输入的是多关键字系统将对用户输入的关键字进行拆分，然后搜索所有含有相关信息的记录返回给用户界面：

（1）用户关键字，并进行提交；

（2）从提交表单中提取数据，并进行相应判断；

（3）连接数据库，建立记录集，用查询语句对表中数据进行查询；

（4）将结果进行加工显示给用户；

（5）结束；

思考题

6-1 什么是认知语言学？

6-2 认知语言学研究的主要内容是什么？

6-3 概述乔姆斯基的四类形式文法。

6-4 什么是扩充转移网络？举例说明它的工作原理。

6-5 什么是认知语义学？主要内容是什么？

6-6 心理空间理论中如何实现概念整合?

思考题

6-7 机器翻译的一般过程包括哪些步骤？试述每个步骤的主要功能是什么？

6-8 请说出神经机器翻译的关键技术。

6-9 脑语言的功能区可分为运动性语言中枢和感觉性语言中枢，请扼要介绍对语言的语义、音韵和拼字的影响。

6-10 试阐述语言加工的三种功能性成分以及它们在大脑中的可能表征。

Thank You