自然语言处理方法与应用

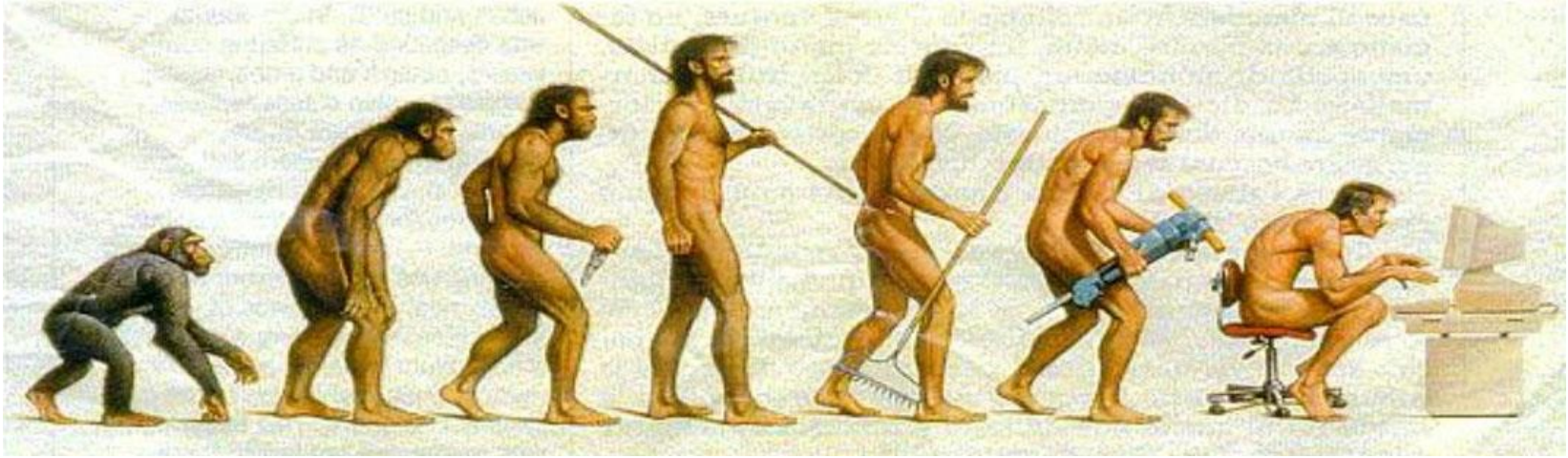
提 纲

1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. 技术现状
6. 我们团队

1. 学科产生与发展

◆什么是自然语言？

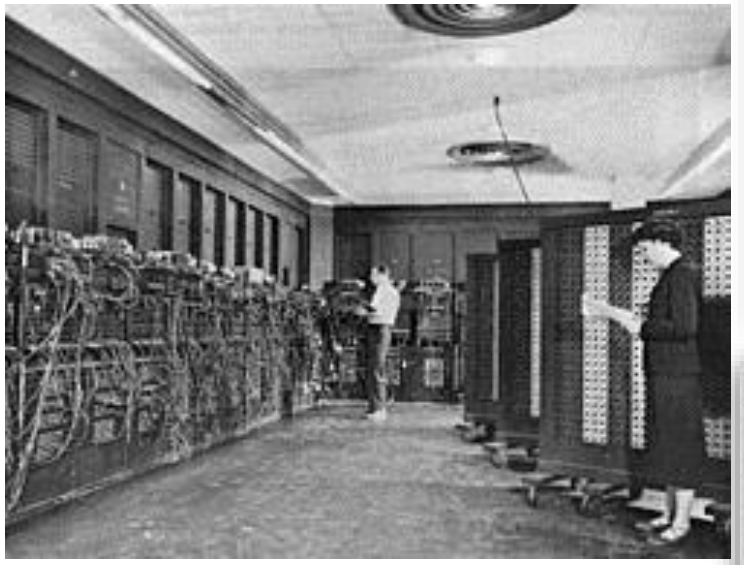
- 自然语言是人类社会发展过程中自然产生的语言，是最能体现人类智慧和文明的产物



- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具
- 人类历史上以语言文字形式记载和流传的知识占知识总量的**80%**以上

1. 学科产生与发展

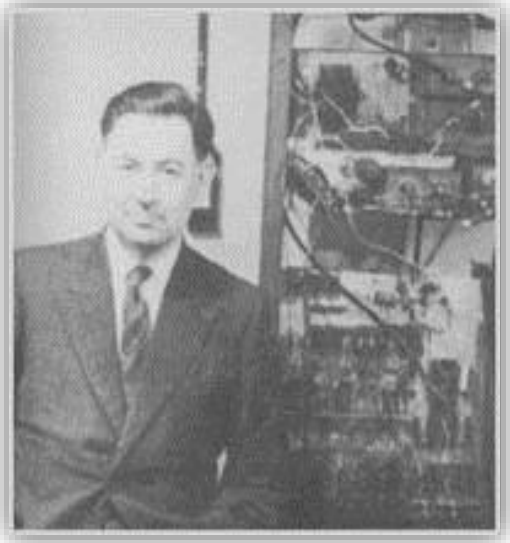
- 自然语言处理技术的诞生



1946年，世界上第一台计算机ENIAC诞生

Warren Weaver

- ✧信息论先驱
- ✧1920至1932年威斯康星大学数学教授
- ✧1932至1955年担任Rockefeller Institute自然科学部主任



A. D. Booth

- ✧数学物理学家
- ✧1947年3月至9月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. 学科产生与发展

诺伯特·维纳 (Norbert Wiener) (1894年11月26日～1964年3月18日)

[Reproduced by permission of the Rockefeller Foundation Archives]

March 4, 1947

Dear Norbert:

I was terribly sorry, when in Cambridge recently, that I got unavoidably held up by several unexpected jobs, and did not get a chance to see you.

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

I wondered if it were unthinkable to design a computer which would translate

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Cordially,

Warren Weaver.

Professor Norbert Wiener
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

WW:AEB

1. 学科产生与发展

达特茅斯
(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)

自然语言理解(natural language understanding, NLU)**成为人工智能研究的核心问题之一。**

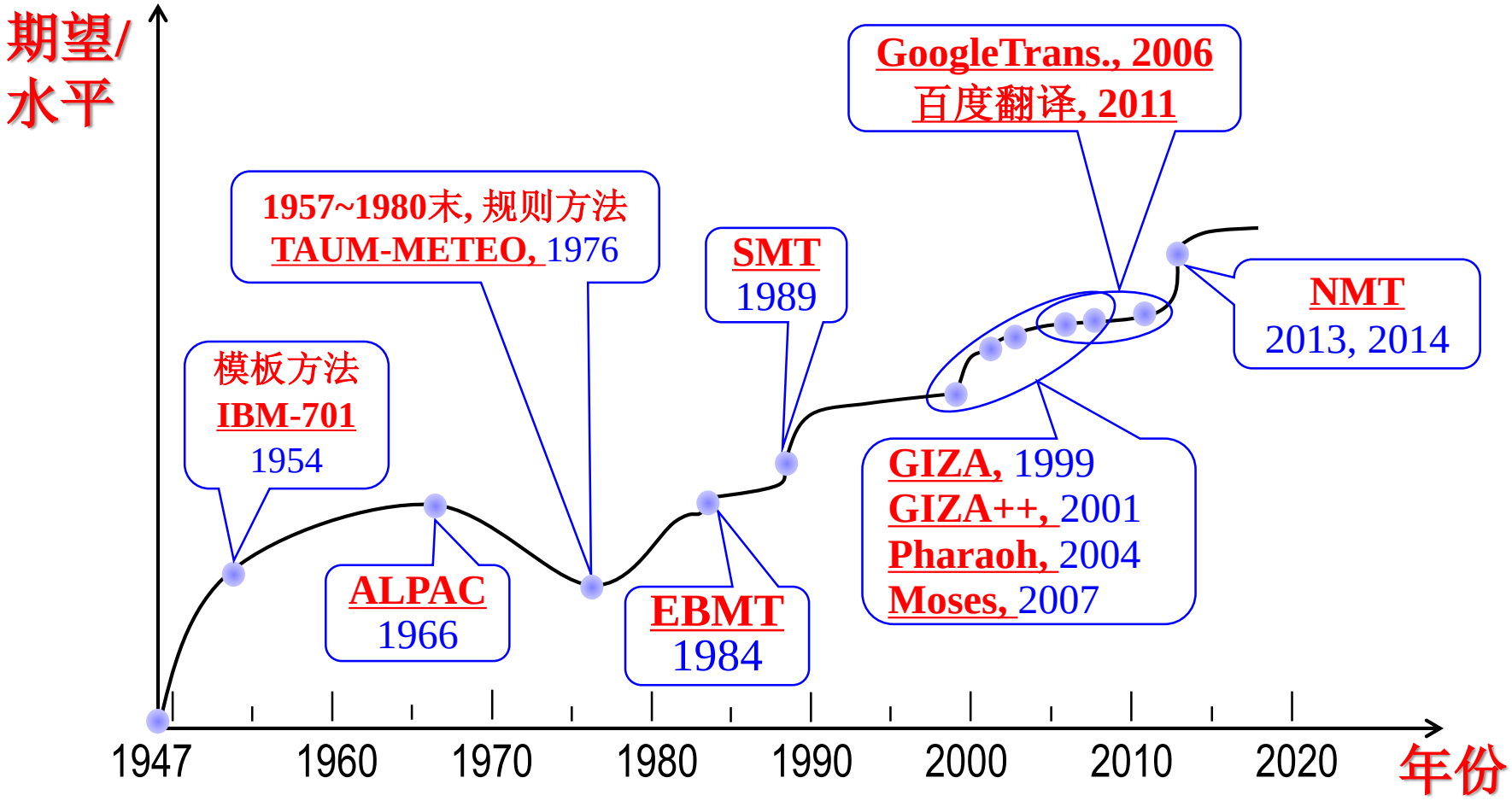
2. 学科产生与发展

- 自W. Weaver 和 A. D. Booth 提出机器翻译概念后，美国和英国的学术界对机器翻译(machine translation, MT)产生了浓厚的兴趣，并得到了实业界的支持。
- 1954年 Georgetown 大学在 IBM 协助下，用IBM-701计算机实现了世界上第一个 MT 系统，实现俄译英翻译，1954年1月该系统在纽约公开演示。系统只有250条俄语词汇，6 条语法规则，可以翻译简单的俄语句子。
- 随后10 多年里，MT研究在国际上出现热潮。

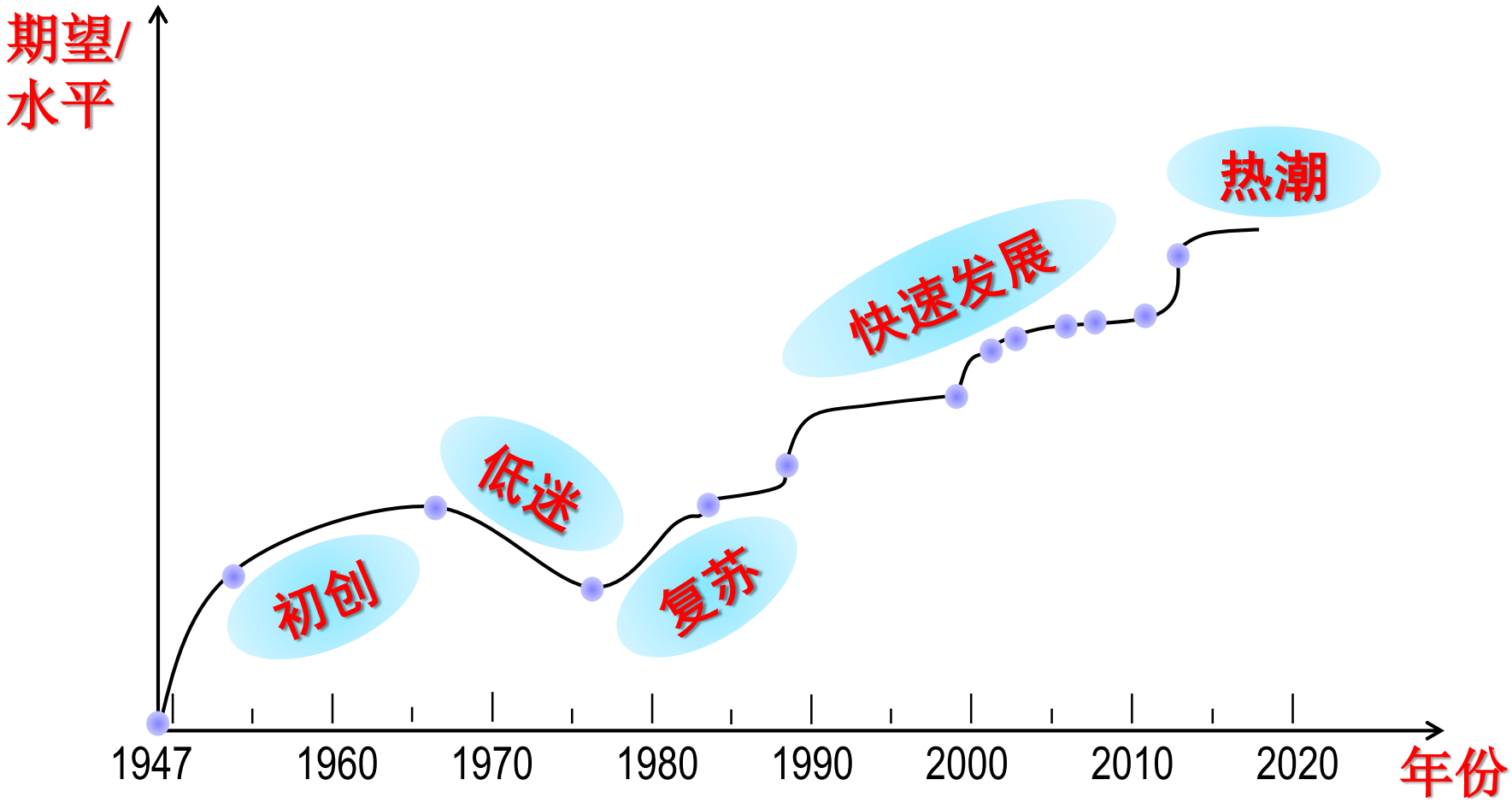
2. 学科产生与发展

- 1962年国际计算语言学学会 (Association for Computational Linguistics, ACL)成立；1965年国际计算语言学委员会 (International Committee on Computational Linguistics, ICCL)成立。
- 1964年，美国科学院成立语言自动处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)，调查机器翻译的研究情况，并于1966年11月公布了一个题为“语言与机器”的调查报告，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”，“机器翻译遇到了难以克服的语义障碍 (semantic barrier)”。从此，机器翻译研究在世界范围内进入低迷状态。计算语言学(computational linguistic, CL)术语首次以正式身份出现在这个报告里。
- 1980S，随着计算机网络的快速发展和普及，以开发实用自然语言处理系统为目标的语言工程技术应运而生，自然语言处理 (natural language processing, NLP) 术语由此诞生。

2. 学科产生与发展


期望/
水平
GoogleTrans., 2006
百度翻译, 2011
1957~1980末, 规则方法
TAUM-METEO, 1976
SMT
1989
NMT
2013, 2014
模板方法
IBM-701
1954
GIZA, 1999
GIZA++, 2001
Pharaoh, 2004
Moses, 2007
ALPAC
1966
EBMT
1984
1947
1960
1970
1980
1990
2000
2010
2020
年份

2. 学科产生与发展


期望/
水平
初创
低迷
复苏
快速发展
热潮
1947
1960
1970
1980
1990
2000
2010
2020
年份

1. 学科产生与发展

◆定义-1：自然语言理解

自然语言理解是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。

《计算机科学技术百科全书》（宗成庆）

清华大学出版社，2018.5

1. 学科产生与发展

◆定义-2：计算语言学

通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科**自然语言处理**相比较，计算语言学更加侧重基础理论和方法的研究。

《计算机科学技术百科全书》（常宝宝）

清华大学出版社，2018.5

1. 学科产生与发展

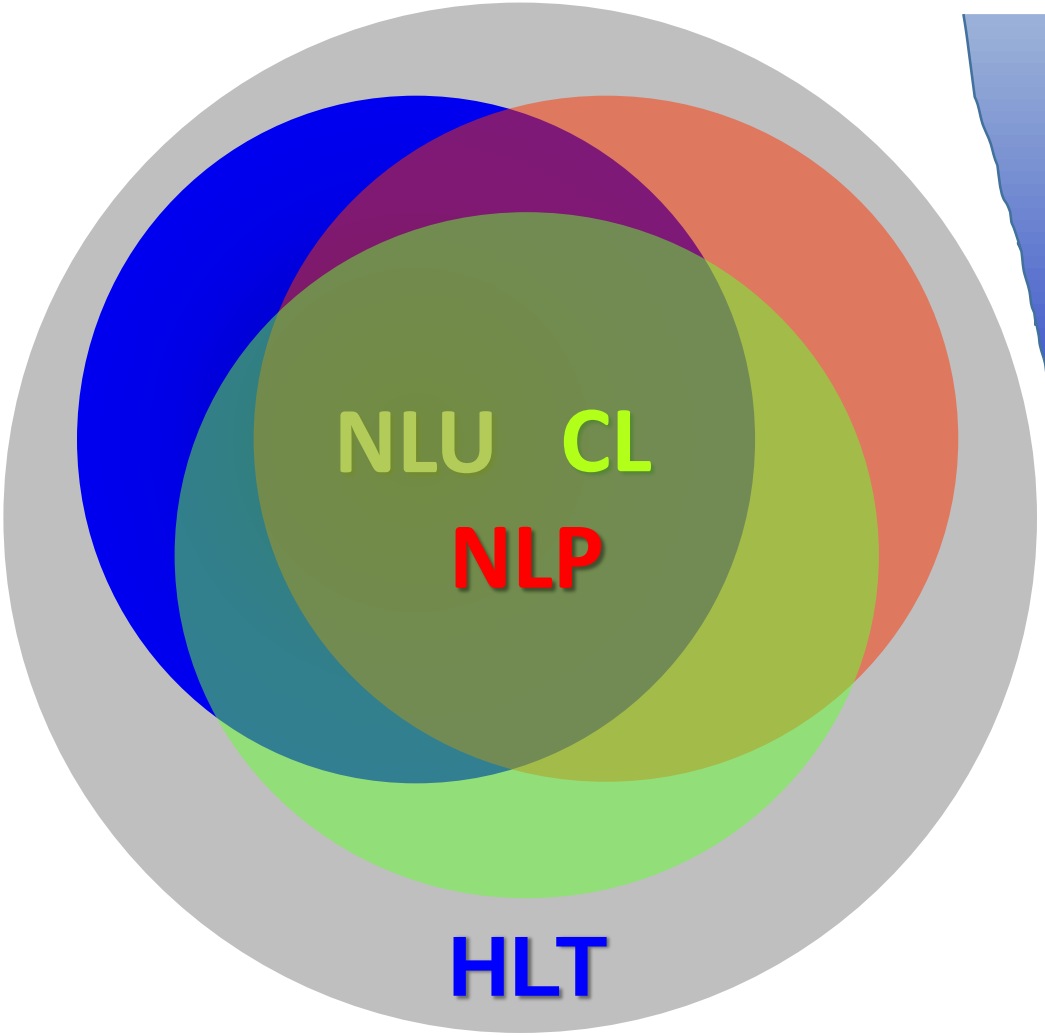
◆定义-3：自然语言处理

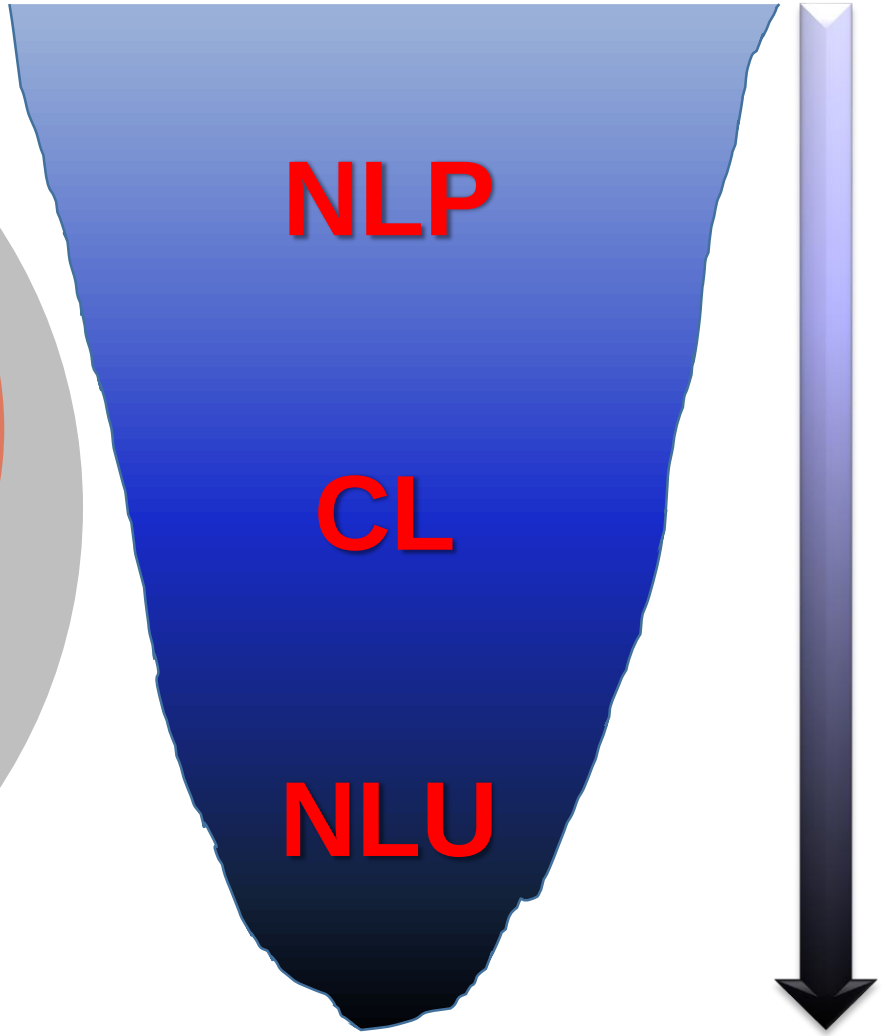
自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

《计算机科学技术百科全书》（宗成庆）

清华大学出版社，2018.5

1. 学科产生与发展





统称：人类语言技术 (human language technique, HLT)

1. 学科产生与发展

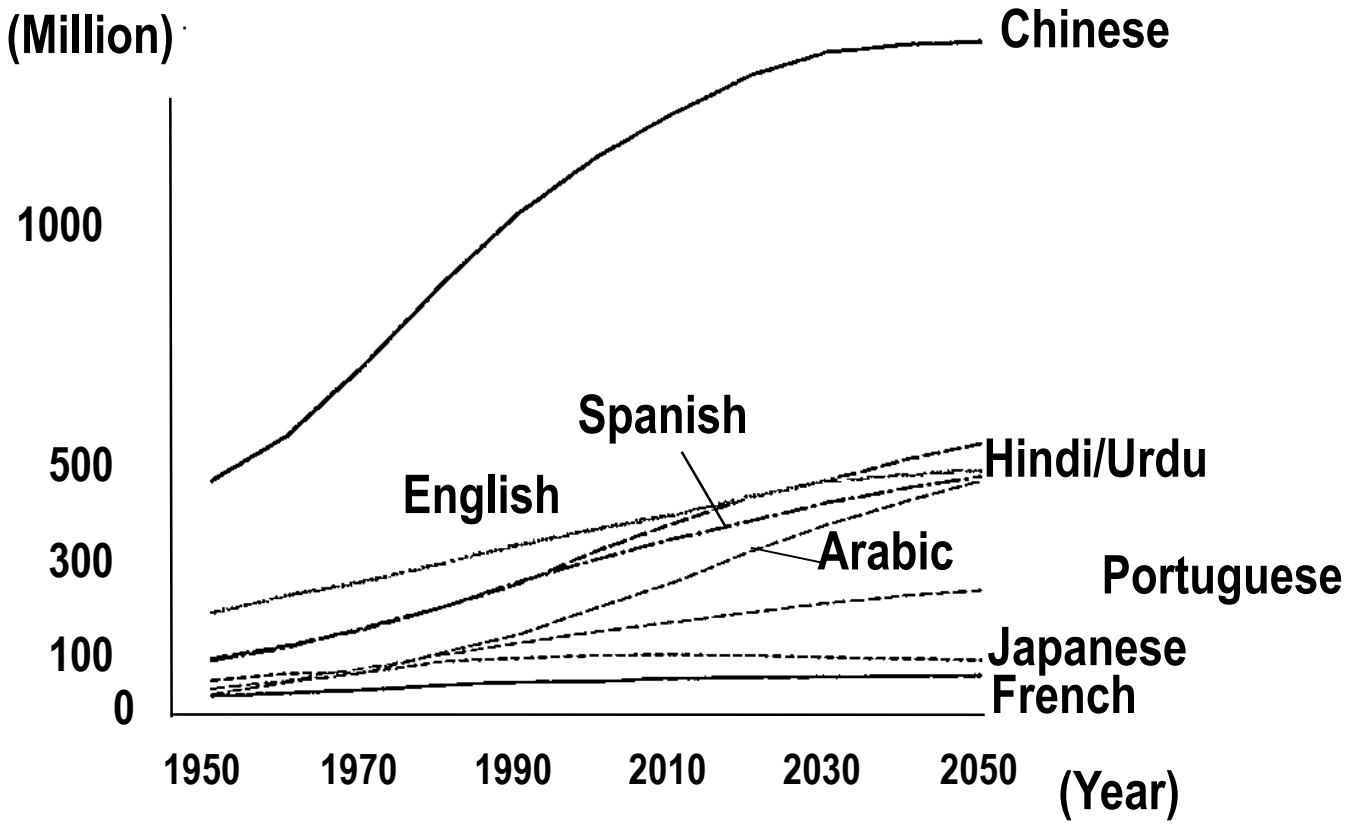
◆定义-4：中文信息处理

以中文为主要处理对象的自然语言处理技术。

在不引起混淆的情况下，通常指汉语。但是从广义上讲，中文指中国境内使用的语言，可以包括汉语、蒙古语、藏语、维吾尔语等几十种语言。

中文信息处理不仅仅是中国人自己感兴趣的研究课题，任何一家国际互联网公司都从来不敢忽略或藐视它。

1. 学科产生与发展


(Million)
1000
500
300
100
0
Chinese
Spanish
English
Hindi/Urdu
Arabic
Portuguese
Japanese
French
1950
1970
1990
2010
2030
2050
(Year)

1. 学科产生与发展

◆ 学科的理论意义

- 探索人脑语言理解的本质，揭示语言认知的奥秘
- 研究和建立计算语言学理论体系
- 推动相关学科产生与发展

◆ 学科的应用价值

- 打破人类语言的障碍，实现任意时间、任意地点、任意语言的无障碍自由通信
- 准确理解人的意图，提高个性化信息服务的质量
- 建立高智能人机交互系统
- 保障网络内容安全，维护国家和公共安全
- 保护民族语言文化，促进全球化社会发展 ……

1. 学科产生与发展

络内容为非结构化文本。

Recommend
火龙果软件

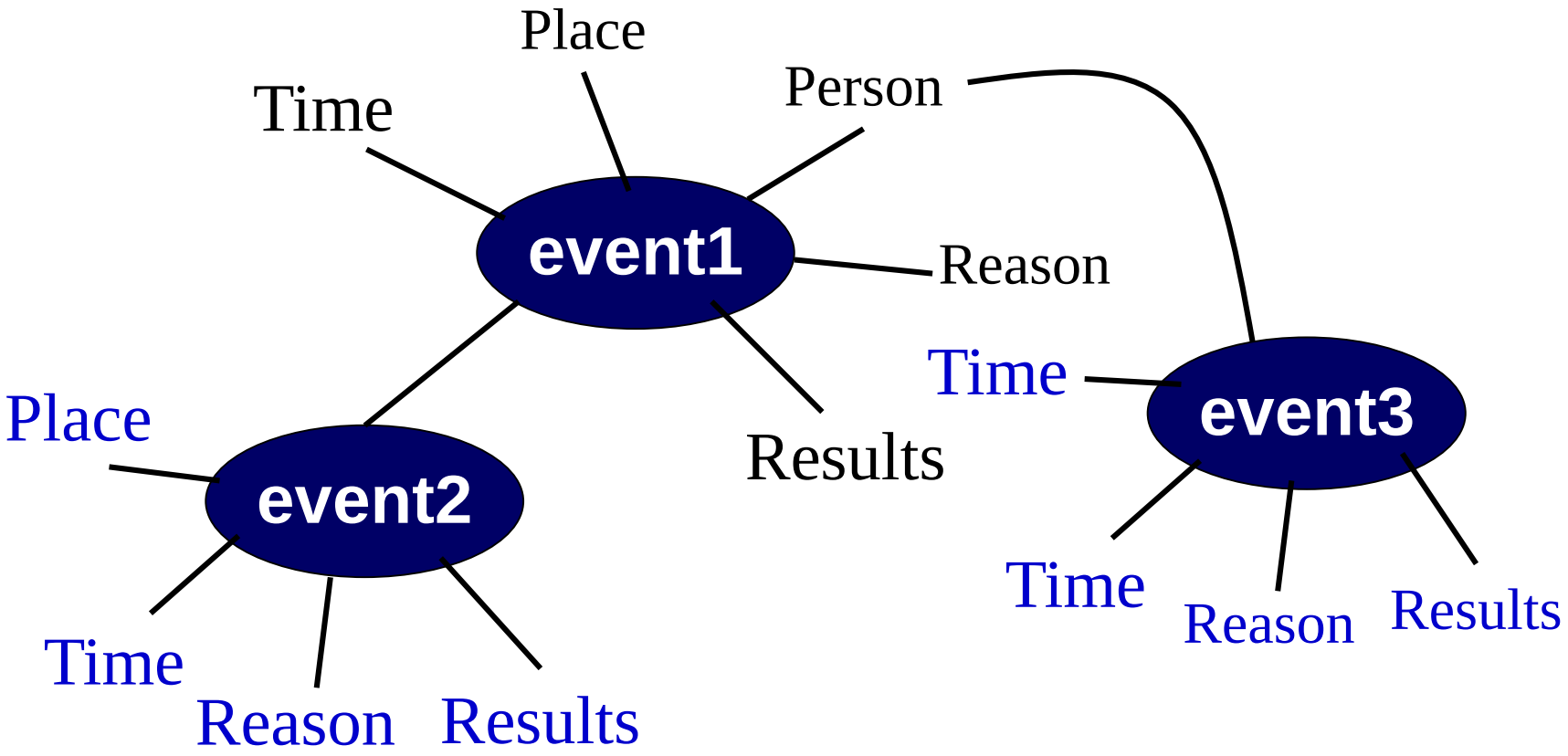
1. 学科产生与发展

络内容为非结构化文本。

1. 学科产生与发展

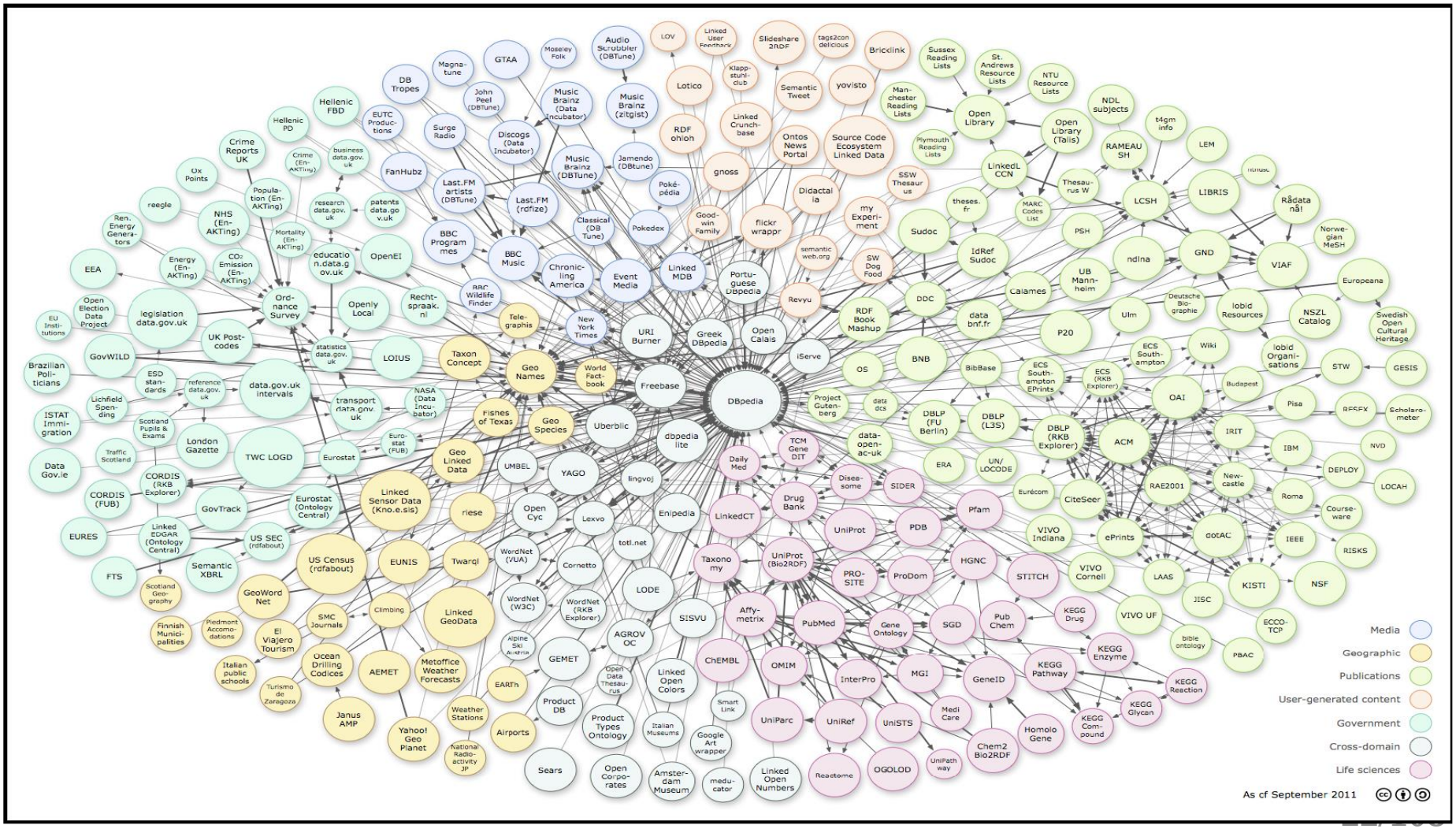
1. 学科产生与发展

人物、事件关系分析意义重大



1. 学科产生与发展

人物、事件关系分析意义重大



1. 学科产生与发展

俄罗斯
西亚
外蒙
朝韩日
印度、巴基斯坦等
东盟
Oroqen
Evenki
Buriat
Daur
Mandarin
Manchu
Korean
Mongolian
Kazakh
Xibe
Kyrgyz
Kalmyk
Uygur
Wakhi
Mandarin
Yugur
Jinyu
Kalmyk
Tu
Salar
Amdo
Choni
Mandarin
Baima
Panang
Tibetan
Qiang
Jiarong
Wu
Huizhou
Gan
Min-Bei
Min-Dong
Pu-Xian
Min-Zhong
Min-Nan
Hakka
Xiang
Hmong
Miao
Nuosu
Lipo
Naxi
Lisu
Nasu
Bouyei
Dong
Zhuang
Yue
(Cantonese)
Ping
Hani
Bugan
Honi
Nisu
Axi
Lolopo
Bai
Xishanba
Tai-Nua
Riang
Parauk
Lu
Hlai
Min-Nan
Luoba
Drung
Groma

1. 学科产生与发展

64个国家和地区
44亿人口
50多种语言

一带一路

出境游人数破亿，
前20个出境游目
的地有12种语言

1. 学科产生与发展

◆说明

语音和文字是语言的两个基本属性，一般将语音识别、语音合成和说话人识别等以语音信号为主要研究对象的语音技术独立出来，而其他以文本(词汇/句子/篇章等)为主要处理对象的研究内容作为自然语言处理的主体。

文字识别更多地涉及图像识别与理解的问题。

在实际应用中，很多技术密切相关，往往需要几种技术集成，共同实现某个应用系统，如人机对话系统、语音翻译系统等。

提 纲

1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. 技术现状
6. 我们团队

2. 技术挑战

- **大量存在的未知语言现象**

  如：高山、高升；吉林、武夷山、桂林、温泉、温馨、时光；
  虎蝇，埃博拉，奥特，闷骚；BoW, word2vec

- **无处不在的歧义词汇**

  如：苹果、粉丝；bank, interest ……
  那辆白色的车是黑车/ 臭豆腐真香啊！

- **复杂或歧义结构比比皆是**

  喜欢乡下的孩子。
  上大学子烛光追思钱伟长。
  “动物保护警察”明年上岗。

2. 技术挑战

- **普遍存在的隐喻表达**

  在微信圈里潜水；打铁还要自身硬；你简直是个木头脑袋！

  雪域高原处处盛开民族团结之花；

  各族人民要像石榴籽一样紧紧拥抱在一起。

- **对翻译而言，不同语言之间的概念不对等**

  馒头： steamed bread

  句子： We do chicken right.

核心挑战：

- 语义概念的表示和计算方法
- 说话人意图的分析和推断
- 语用场景的分析

2. 技术挑战

ICML, 6-11 July 2015, Lille, France
Panel Discussion
Deep Learning Workshop ICML'15

2. 技术挑战

At DL 2015, Neil Lawrence said …

“NLP is kind of like a rabbit in the headlights of the deep learning machine, waiting to be flattened.”

A Professor of Machine Learning at the University of Sheffield

提 纲

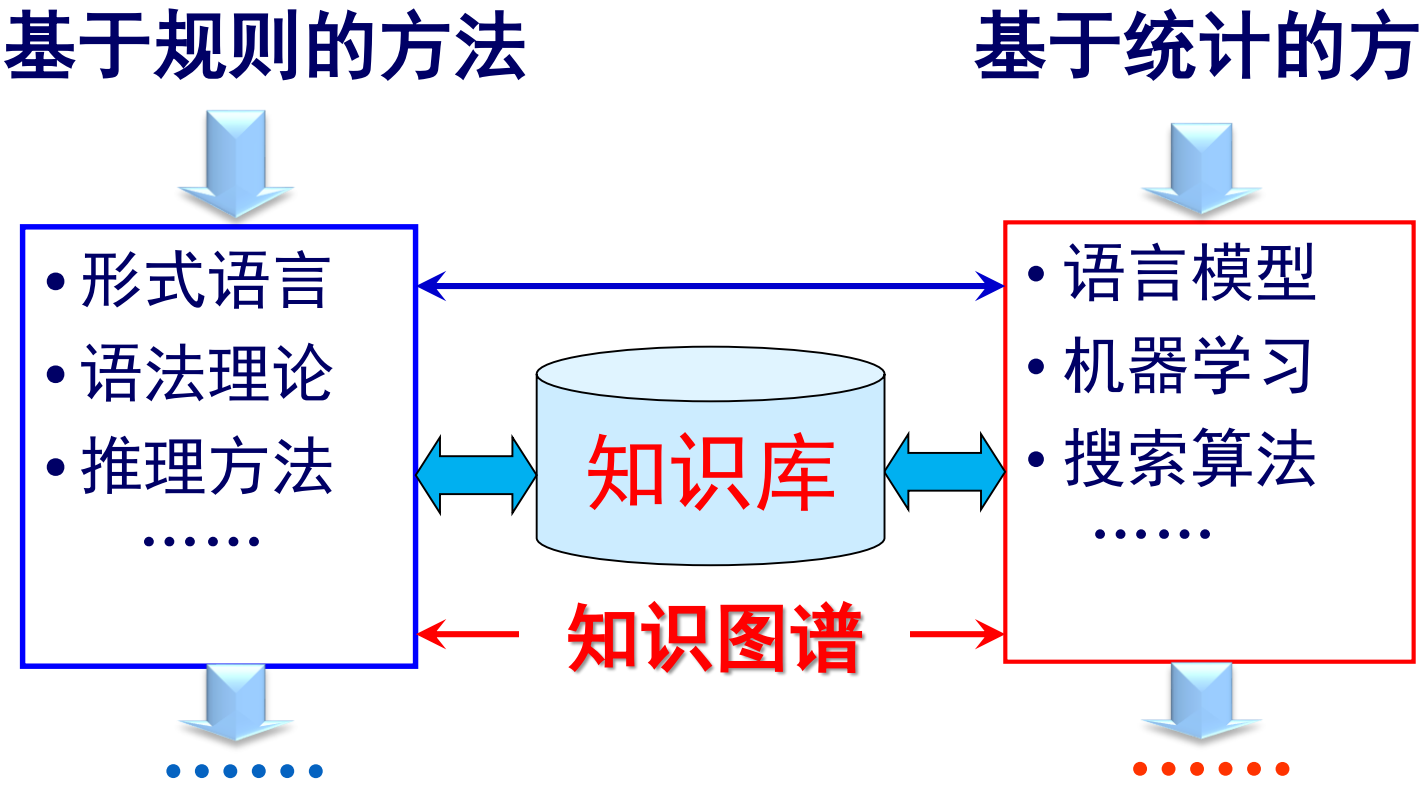
1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. 技术现状
6. 我们团队

3. 基本方法

4.1 方法概述

- **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（符号逻辑）**
- **经验主义方法**：～1950S，1980S～
  - 大规模训练样本
  - 数据驱动的统计模型**—基于统计的方法（统计学习）**

3. 基本方法



理性主义与经验主义的合谋 —
符号智能 + 计算智能，建立融合方法

3. 基本方法

4.2 传统的统计学习方法 (1980s~)

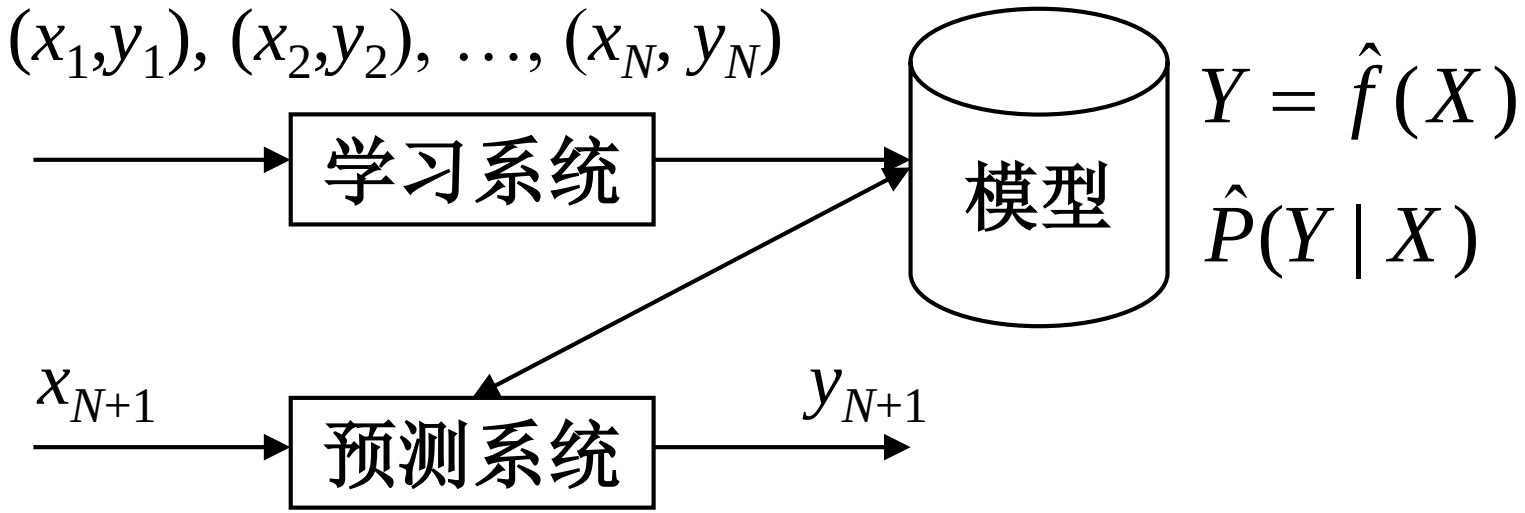
- ✧ 语音识别，如 SPHINX 语音识别系统(CMU)
- ✧ 训练控制系统用于驾驶车辆，如 ALVINN 系统
- ✧ 在各种大规模数据库中发现隐藏的一般规律，如美国国家航空和航天局(NASA)使用决策树进行天体分类
- ✧ 世界级水平的西洋双陆棋博弈

学习：通过经验提高性能

统计学、信息论、计算复杂性理论、人工智能、神经生物学等相关学科的发展进一步推动了机器学习研究

3. 基本方法

基本框架：



给定一个训练数据集：$T=\{(x_1,y_1), (x_2,y_2), \ldots, (x_N, y_N)\}$，其中，$(x_i,y_i)$，$i=1, 2, \ldots, N$，称为样本。$x_i$是输入的观测值，也称输入或实例；$y_i$ 是输出的观测值，也称输出。

3. 基本方法

4.3 常用的统计模型和开源工具

◆统计模型（生成式+区分式）

- 语言模型 (language model)
- 隐马尔可夫模型(hidden Markov model, HMM)
- *k*-近邻法(*k*-nearest neighbor, *k*-NN)：多类分类问题
- 朴素贝叶斯法(na ïve Bayes)：多类分类问题
- 决策树(decision tree)：多类分类问题
- 最大熵(maximum entropy)：多类分类问题
- 感知机(perceptron)：二类分类
- 支持向量机(support vector machine, SVM)：二类分类
- 条件随机场(conditional random field, CRF)：序列标注

3. 基本方法

◆ 开源工具：

- 语言模型
  - ✧ **SRI** 语言模型工具：
    http://www.speech.sri.com/projects/srilm/
  - ✧ CMU-Cambridge 语言模型工具：
    http://mi.eng.cam.ac.uk/~prc14/toolkit.html
- 隐马尔可夫模型：http://htk.eng.cam.ac.uk/

3. 基本方法

- 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/

3. 基本方法

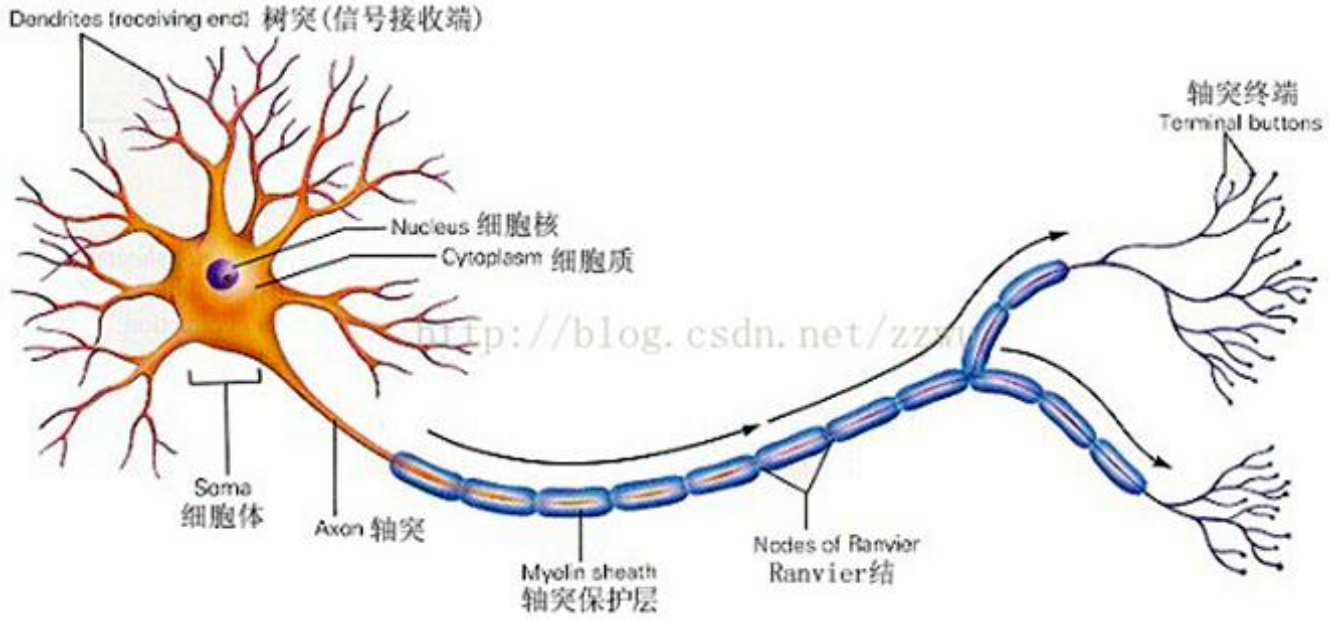
- 最大熵：
  - ✧ **OpenNLP**： http://incubator.apache.org/opennlp/
  - ✧ **Malouf**： http://tadm.sourceforge.net/
  - ✧ **Tsujii**： http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/
  - ✧ 张乐： http://homepages.inf.ed.ac.uk/lzhang10/maxent.html
  - ✧ 林德康： http://webdocs.cs.ualberta.ca/~lindek/downloads.htm
- 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

3. 基本方法

4.4 深度学习方法

- 人工神经网络 (artificial neural networks, ANN)

  1943年心理学家 沃伦•麦卡洛克(W. McCulloch) 和数理逻辑学家 W. Pitts 建立了神经网络的数学模型，提出了神经元的形式化数学描述和网络结构方法



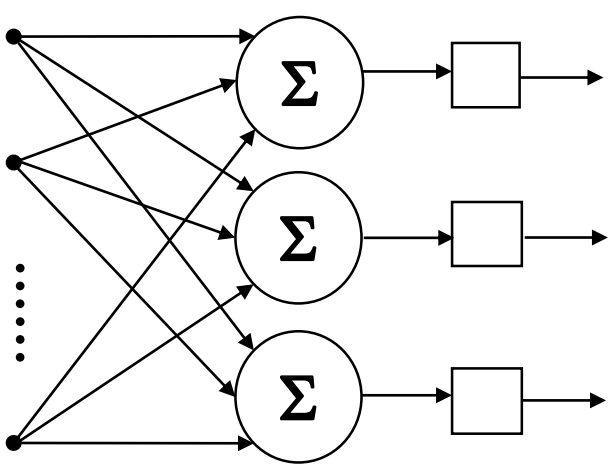
3. 基本方法

- 神经网络 (neural network)

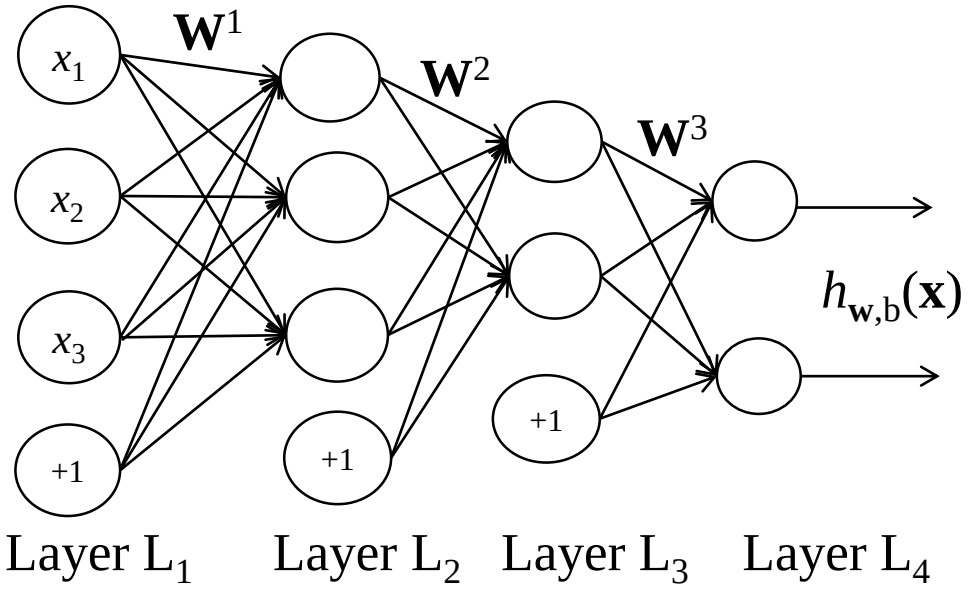
  1982, Hopfield 提出神经网络模型;

  1984, 建立连续时间的Hopfield神经网络模型。

  - 有限个神经元
  - 所有神经元的输入都是同一个向量 $\vec{X}$
  - 网络输出也是一个向量，维数等于神经元的个数

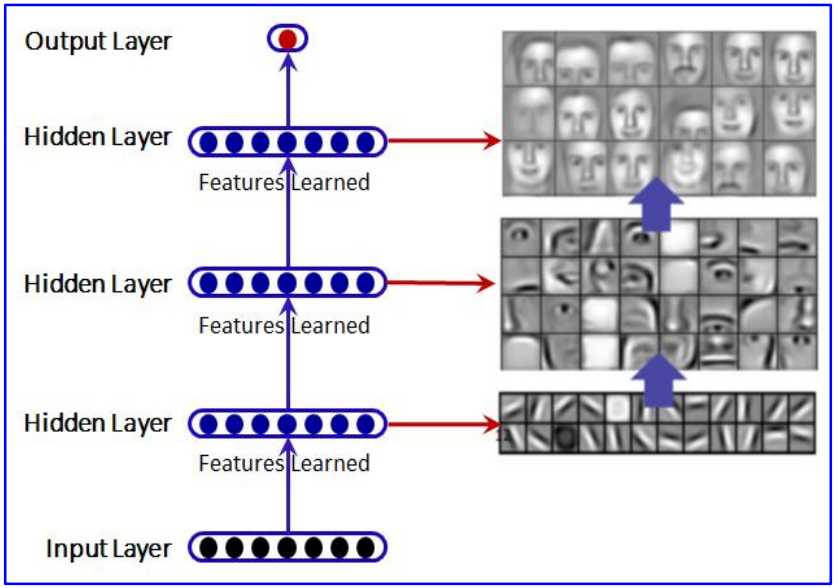


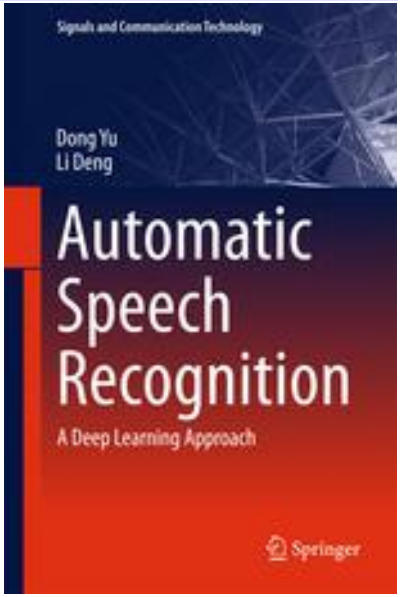
单层网络



3. 基本方法

- 深度学习 (deep learning, DL)
  - 基于深层(前向多层)神经网络的学习通过校正训练样本，对各个层的权重进行调整(learning)
  - 2006 年 G. E. Hinton (辛顿)等人使用受限玻尔兹曼机(restricted Bolzman machine)进行逐层无监督训练方法，率先在图像识别上获得了突破。
  - 2009年DNN在语音识别中获得成功应用





3. 基本方法

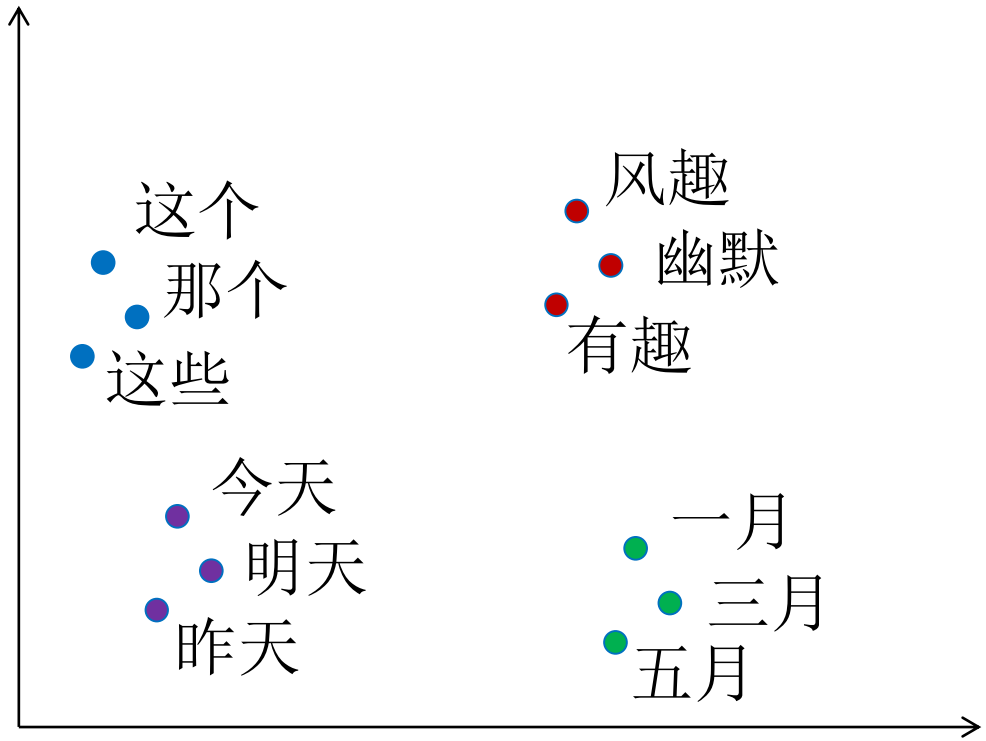
◆ 神经语言模型

在传统的 n 元语法中，$p(s)=\prod_{i=1}^{m+1} p(w_i \mid w_{i-n+1}^{i-1})$

$$p(w_i|w_{i-1}\cdots w_{i-n+1}) = \frac{count(w_i w_{i-1}\cdots w_{i-n+1})}{count(w_{i-1}\cdots w_{i-n+1})}$$

3. 基本方法

◆分布式表示 — 词向量表示 (word2vec)



风趣

$$\begin{bmatrix} 0.24 \\ 0.15 \\ 0.42 \\ 0.51 \\ 0.21 \end{bmatrix}$$

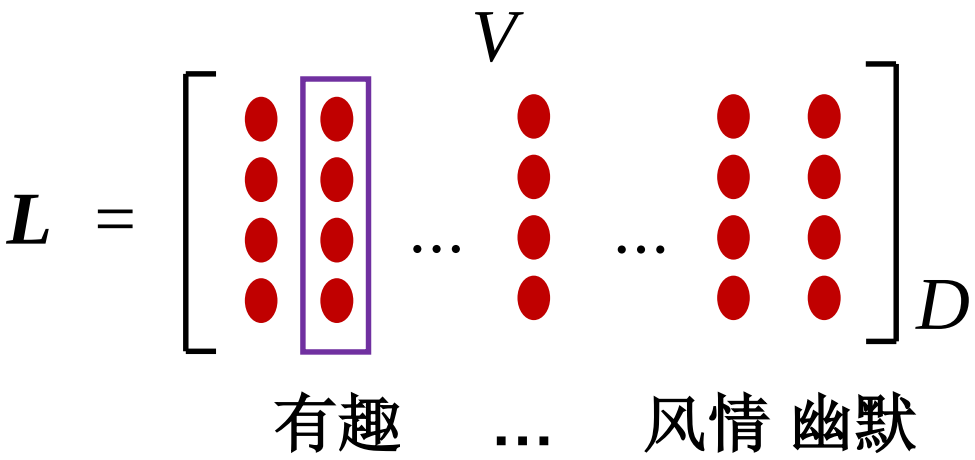
幽默

$$\begin{bmatrix} 0.25 \\ 0.12 \\ 0.39 \\ 0.46 \\ 0.26 \end{bmatrix}$$

低维、稠密的连续实数空间

3. 基本方法

◆基于文本的词向量学习



$L \in R^{D \times V}$

➢词表规模V的确定：

1)训练数据中所有词；

2)频率高于某个阈值的所有词；

3)前V个频率最高的词，e.g. V=50000, V=80000

3. 基本方法

➢开源工具

- NNlm: 前馈神经网络语言模型 (feed-forward n-gram neural language model) http://nlg.isi.edu/software/nplm/
- RNNlm: 循环神经网络语言模型 (recurrent neural language model) http://rnnlm.org/
- LSTMlm: LSTM语言模型 (recurrent neural language model with LSTM unit) https://www-i6.informatik.rwth-aachen.de/web/Software/rwthlm.php
- LSTM 反向传播算法: http://arunmallya.github.io/writeups/nn/lstm/index.html#/
- Google Word2Vec: http://code.google.com/p/word2vec/

… …

提 纲

1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. 技术现状
6. 我们团队

4. 应用举例

5.1 基础问题：汉语自动分词 (Chinese word segmentation)

词是自然语言中具有独立含意的最小单位；汉语属于孤立语，孤立语与部分黏着语（如日语、朝鲜语等）词语之间没有间隔；汉语的语义极其丰富，字和词之间的界限不是非常明确，使用灵活。因此，汉语分词成为自然语言处理中基本的具有挑战性的问题。

例如：

①自动化研究所取得的成就

自动化/研究所/取得/的/成就

自动化/研究/所/取得/的/成就

②门把手弄坏了

门/把/手/弄/坏/了

门把手/弄/坏/了

4. 应用举例

已有的方法：

- 全切分方法
- 最短路径切分方法
- 基于 *n*-gram 的统计方法
- 基于 HMM 的分词与词性标注一体化方法

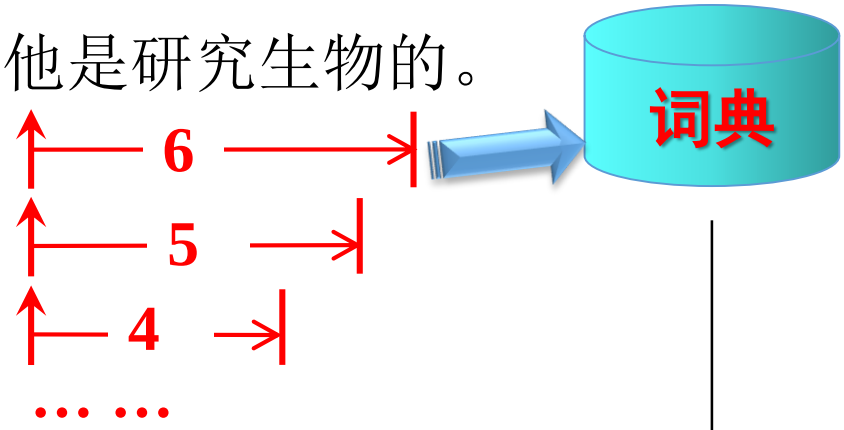
……

几十种之多！

4. 应用举例

(1) 最大匹配法(基于规则或模板的方法)

他是研究生物的。



他/ 是研究生物的。

他/ 是/ 研究生物的。

他/ 是/ 研究生/ 物/ 的/ 。

正向最大匹配法
(forward maximum matching, FMM)

他是研究生物的。 ?

逆向最大匹配法
(backward maximum matching, BMM)

双向最大匹配法
(bi-directional MM)

4. 应用举例

(2) 基于*n*-gram的分词方法(传统的统计方法，生成式)

对于待切分的句子$S = z_1 z_2 \ldots z_m$, 假设 $W = w_1 w_2 \ldots w_k$ $(1 \le k \le n)$ 是一种可能的切分结果。那么，

$$\begin{aligned}\hat{W} &= \arg\max_W p(W \mid S) \\ &= \arg\max_W p(W) \times p(S \mid W) \\ &\cong \arg\max_W p(W)\end{aligned}$$

最基本的做法是以词为独立的统计基元，但效果不佳。

4. 应用举例

(3) 由字构词的分词方法(传统的统计方法，区分式)

(Character-based tagging)[Xue and Converse, 2002]

基本思想：将分词过程看作是字的分类问题，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。通常情况下，只有4种可能的词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

例：上海计划到本世纪末实现人均国内生产总值五千美元。

上/B 海/E 计/B 划/E 到/S 本/S 世/B 纪/E 末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总/B 值/E 五/B 千/M 美/M 元/E 。/S

上海/ 计划/ 到/ 本/ 世纪/ 末/ 实现/ 人均/ 国内/ 生产/ 总值/ 五千美元/ 。/

N. Xue and S. Converse. 2002. Combining Classifiers for Chinese Word Segmentation. *Proc. of the 1st SIGHAN Workshop*, Taipei

4. 应用举例

上/B 海/E 计/B 划/E 到 本 世 纪 ……

B, E, M, S ?

- 当前字的前后 n 个字 (如 $n=\pm 2$)
- 当前字左边字的标记
- 当前字在词中的位置

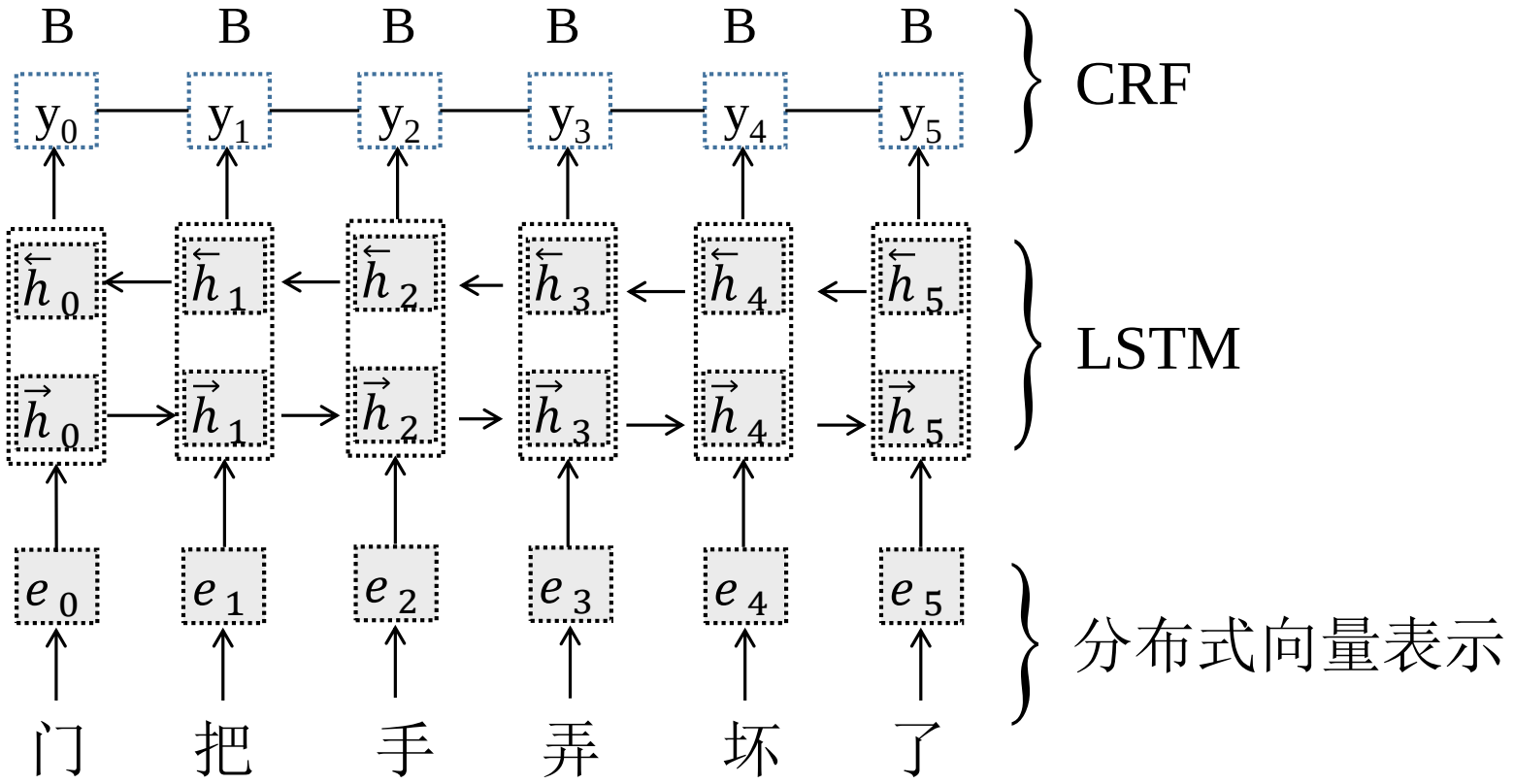
……

- **特征选择**
- **分类器**
  - ✓ 条件随机场 (CRFs)
  - ✓ 支持向量机 (SVM)
  - ✓ 最大熵 (ME)
  - ✓ 贝叶斯

……

4. 应用举例

(4) 基于神经网络的分词方法



4. 应用举例

5.2 应用技术: 机器翻译 (machine translation, MT)

- 基于模板的直接转换法
- 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
  - 基于事例的翻译方法
  - 统计翻译方法
  - 神经网络机器翻译

数据/语料驱动的方法 (Data/Corpus driven)

4. 应用举例

◆基于模板的直接转换法 (Template-based Direct Translation)

从源语言句子的表层出发，将单词、短语或句子直接置换成目标语言译文，必要时进行简单的词序调整。对原文句子的分析仅满足于特定译文生成的需要。这类翻译系统一般针对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。例如：

I like Mary. → 我喜欢玛丽。

X like Y → X 喜欢 Y。

(1) 今天我想吃面包/ 今天我想吃食堂/ 今天我想吃大碗

Today I would like to eat bread. ✓ canteen / big bowl ✗

(2) 学英语/ 学钢琴

study English/ ✓ study piano ✗

(3) 写文章/ 写黑板/ 写大仿 ？

4. 应用举例

- **基于规则的翻译方法** (Rule-based MT, RBMT)

1957年美国学者V. Yingve在《句法翻译框架》(Framework for Syntactic Translation) 一文中提出了对源语言和目标语言均进行适当描述、把翻译机制与语法分开、用规则描述语法的实现思想，这就是基于规则的翻译方法。

基于规则的翻译过程分成6个步骤：

(*a*) 对源语言句子进行词法分析

(*b*) 对源语言句子进行句法/语义分析

(*c*) 源语言句子结构到译文结构的转换

(*d*) 译文句法结构生成

(*e*) 源语言词汇到译文词汇的转换

(*f*) 译文词法选择与生成

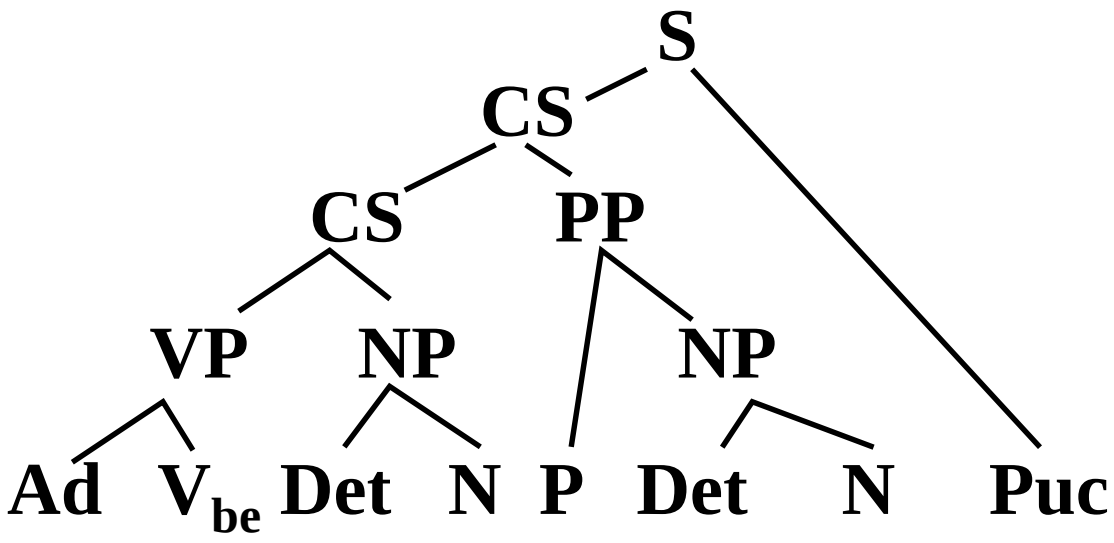
4. 应用举例

给定源语言句子：**There is a book on the desk.**

(*a*)词法分析：

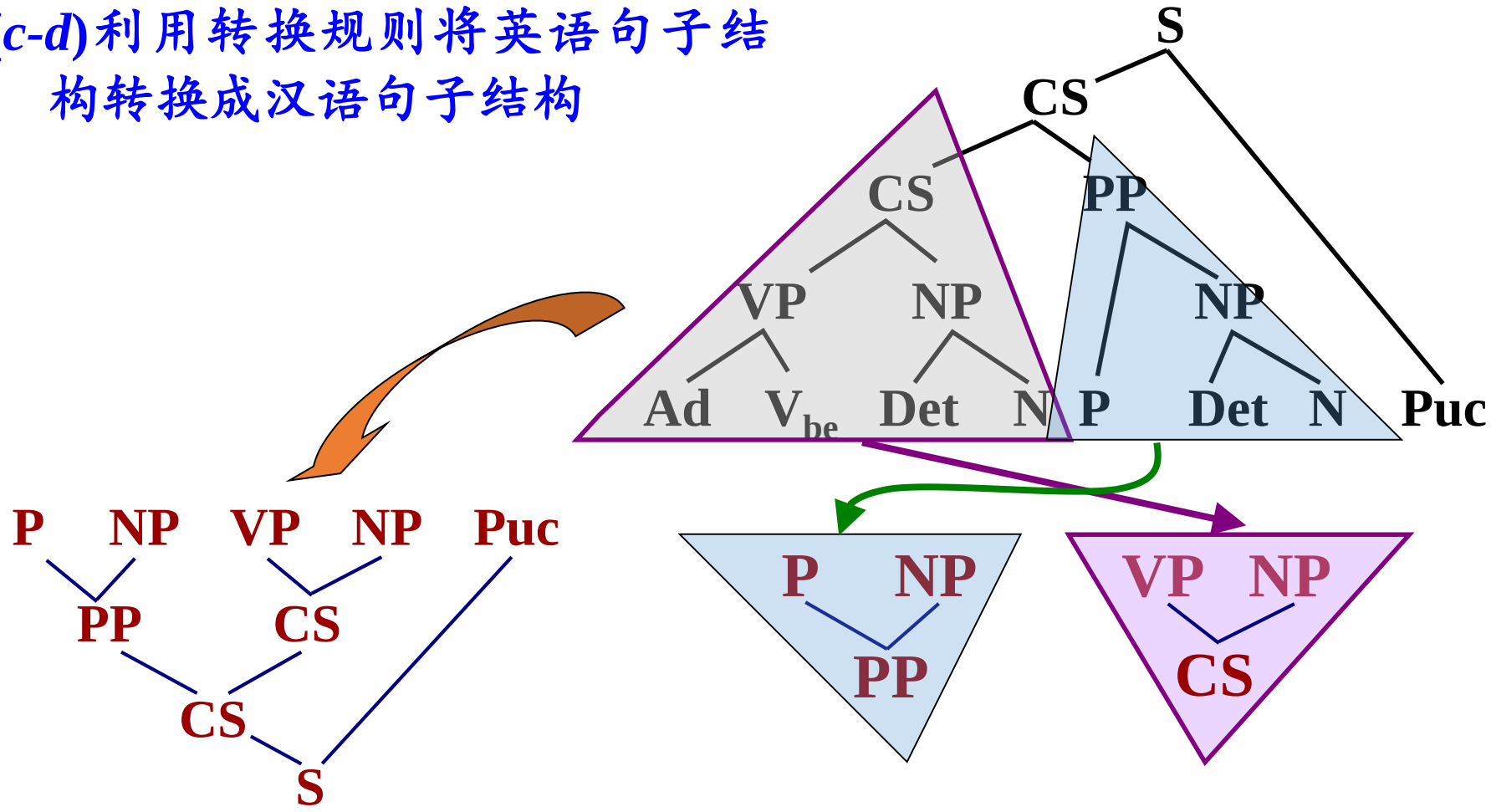
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

(*b*)利用句法规则进行句法结构分析：



4. 应用举例

(c-d)利用转换规则将英语句子结构转换成汉语句子结构



4. 应用举例

*(e-f)*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

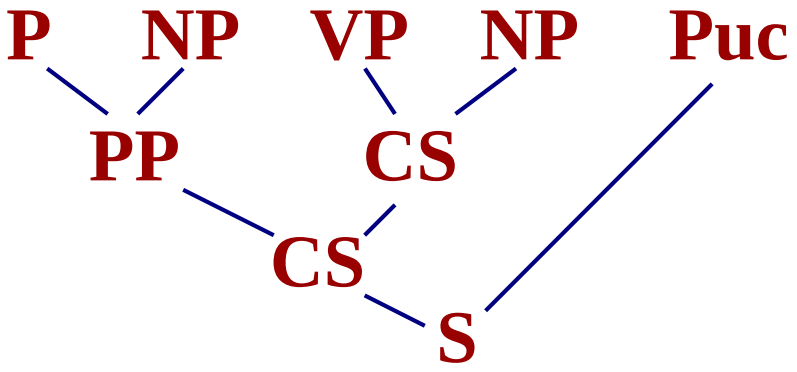
#a, Det, 一

#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。

由于基于规则的翻译方法执行过程为“独立分析—独立生成—相关转换”， 因此，又称基于转换的翻译方法。

4. 应用举例

- 规则方法评价：

优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

弱点：规则一般由人工编写，工作量大，主观性强，一致性难以保障，不利于系统扩充，对非规范语言现象缺乏相应的处理能力。

4. 应用举例

- **统计机器翻译方法** (Statistical MT, MT)
  - 基于1947年 W. Weaver 提出的“解读密码”思想。
  - 1990年IBM的Peter F. Brown 等在*Computational Linguistics* 上发表的论文“统计机器翻译方法”[Brown et al., 1990]和1993年他们发表在该杂志的“统计机器翻译的数学：参数估计”[Brown et al., 1993]两篇文章奠定了统计机器翻译的理论基础。
  - 基于统计方法实现的*Candide* 翻译系统[Berger et al.,1994]在ARPA组织的机器翻译评测中表现良好。

4. 应用举例

- 双语平行句对

merkezdiki dölet apparatliri bilen jaylardiki dölet apparatlirining xizmet hoquqi merkezning bir tutash rehberlikide jaylarning teshebbuskarliqi we aktipliqini toluq jari qildurush prinsipi boyiche ayrilidu.

中央和地方的国家机构职权的划分，遵循在中央的统一领导下，充分发挥地方的主动性、积极性的原则。

madda jungxua xelq jumhuriyitide hemme millet bapbarawer.

中华人民共和国各民族一律平等。

herqandaq milletni kemsitish we ëzishni men'i qilidu, milletler ittipaqliqini buzidighan we milliy bölgünchilik qilidighan qilmishlarni men'i qilidu.

禁止对任何民族的歧视和压迫，禁止破坏民族团结和制造民族分裂的行为。 ……

4. 应用举例

- 数学描述

源语言句子： $S = s_1^m \equiv s_1 s_2 \cdots s_m$

目标语言句子： $T = t_1^l \equiv t_1 t_2 \cdots t_l$

$$p(T|S) = \frac{p(T) \times p(S|T)}{p(S)}$$

$$\hat{T} = \arg\max_T p(T) \times p(S|T)$$

语言模型
Language model, LM

翻译模型
Translation model, TM

4. 应用举例

实现搜索过程的模块称为解码器(decoder)。

统计翻译中的三个关键问题：

(1) 估计语言模型概率 $p(T)$;

(2) 估计翻译概率 $p(S|T)$;

(3)快速有效地搜索 T 使得 $p(T) \times p(S \mid T)$ 最大。

4. 应用举例

- **神经机器翻译方法**

  基于神经网络的机器翻译 (Neural Network based MT, NNMT)
  或称神经机器翻译(Neural Machine Translation, NMT)，2013

Yoshua Bengio
Montreal Institute for Learning Algorithms (MILA)

Kyunghyun Cho
New York University (NYU)

4. 应用举例

- 基本原理

给定源语言句子： $C = c_1^l \equiv c_1 c_2 \cdots c_l$

将其翻译成目标语言句子： $E = e_1^m \equiv e_1 e_2 \cdots e_m$

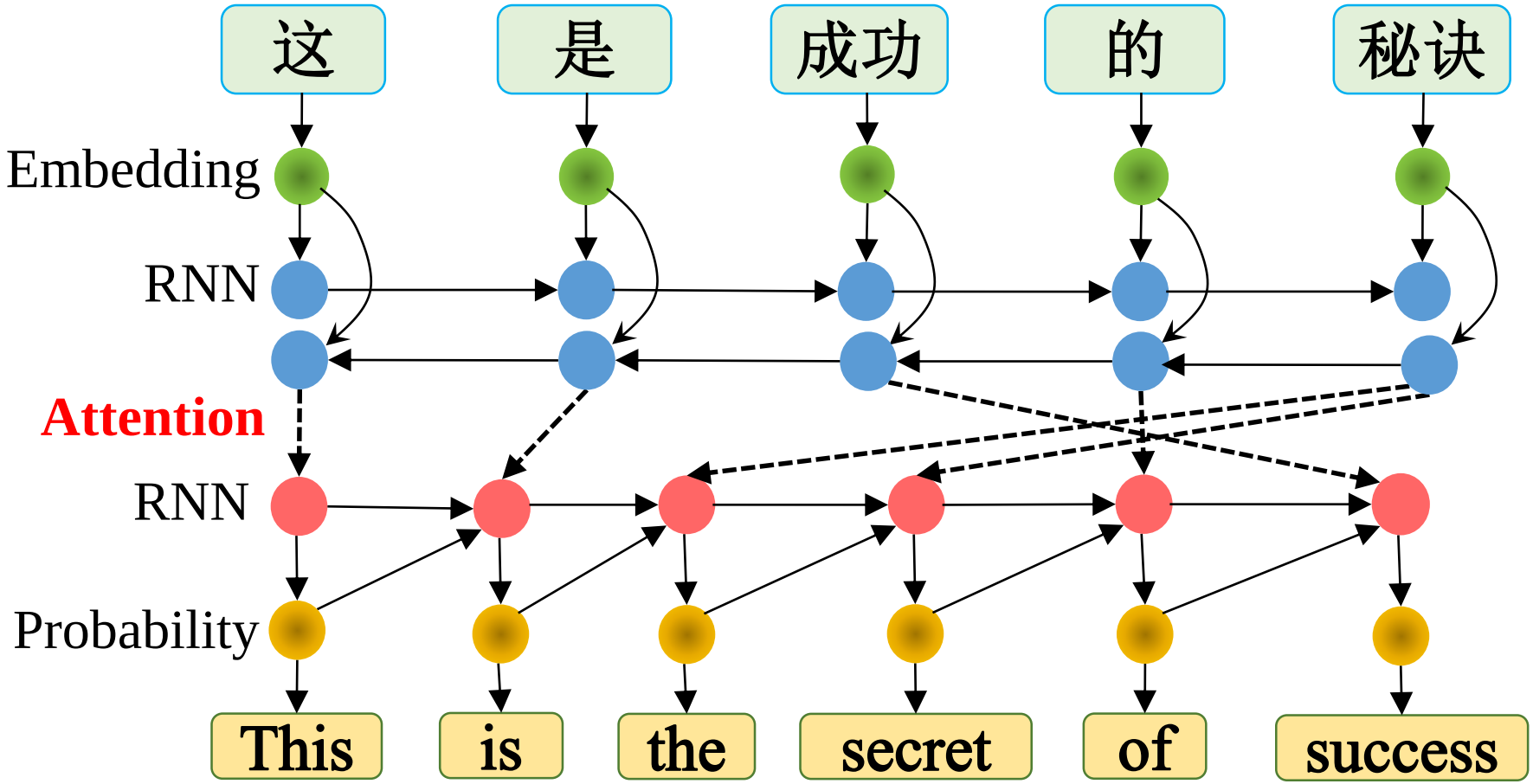
词汇向量化表示：Word2Vec

$$P(e_i) \approx P(e_i | e_1 \cdots e_{i-1}, C)$$

目标函数： $$L = \sum_i log\big(P(e_i|C)\big)$$

4. 应用举例

$$h_t = \tanh(W^* x + U^* h_{t-1} + b)$$



Transformer 的核心是注意机制，包括自注意力(self-attention)和跨语言注意力(cross-language-attention)。

4. 应用举例

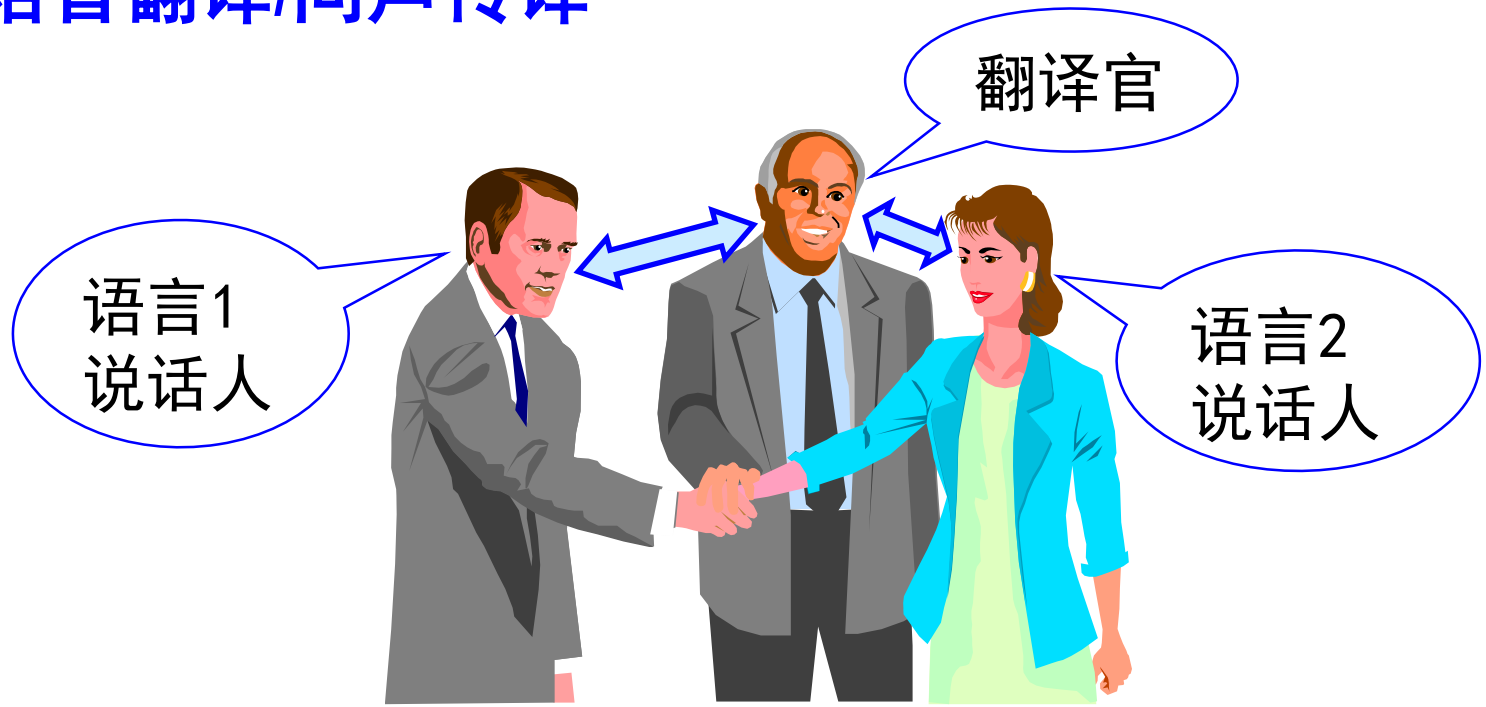
- 数据驱动的方法评价：

优点：一般不需要对源语言句子进行深层次分析，甚至对源语言没有任何基本知识，只要有足够多的双语言句对就可以建立一个机器翻译系统。

弱点：对于某些语言对来说，收集高质量大规模的双语句对并不是一件容易的事情；机器翻译过程和译文结果缺乏解释性；对于复杂句子、生僻词汇、指代等问题仍缺乏得利的处理手段。

4. 应用举例

5.3 语音翻译/同声传译



源语言(Source language) 1

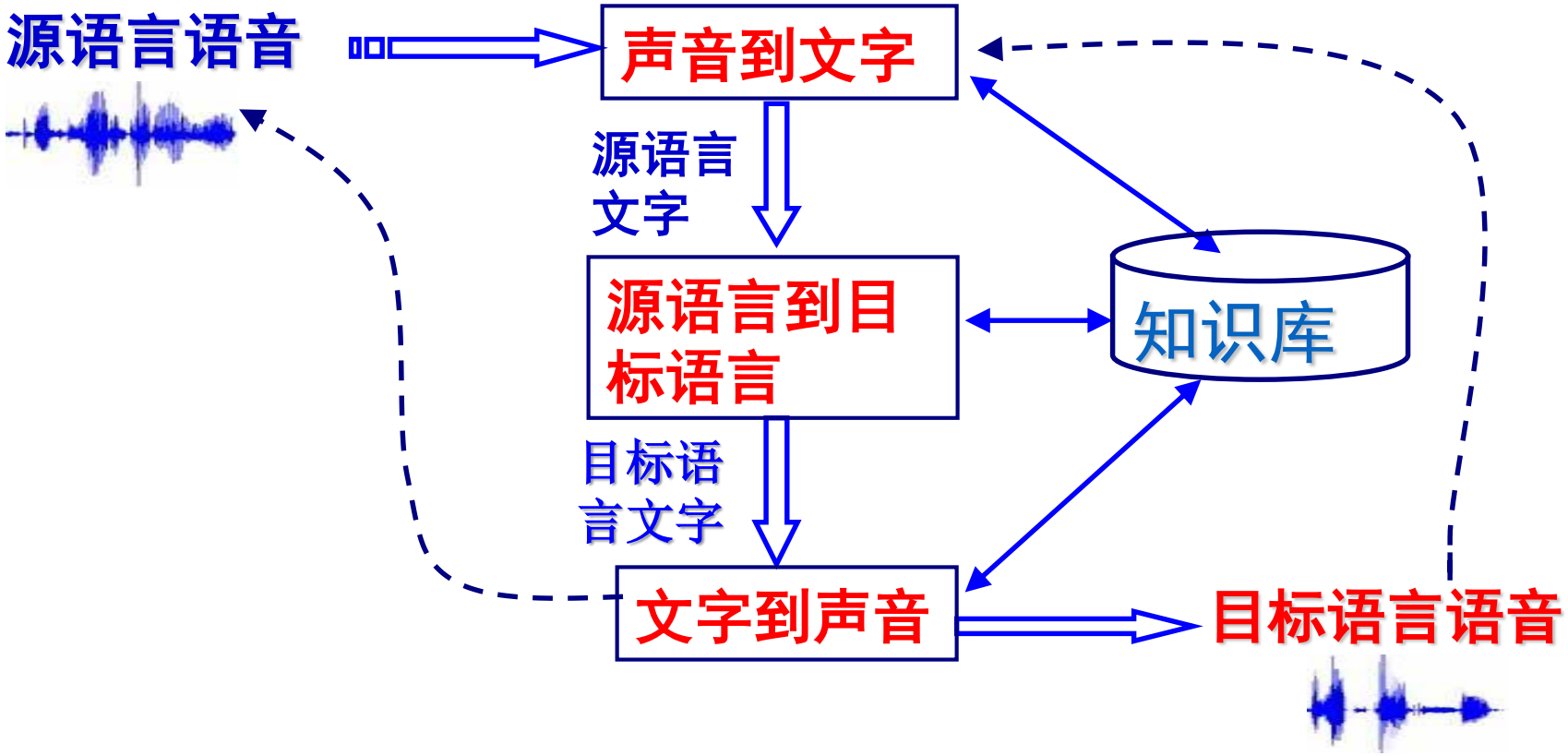
源语言(Source language) 2

翻译器
Translator

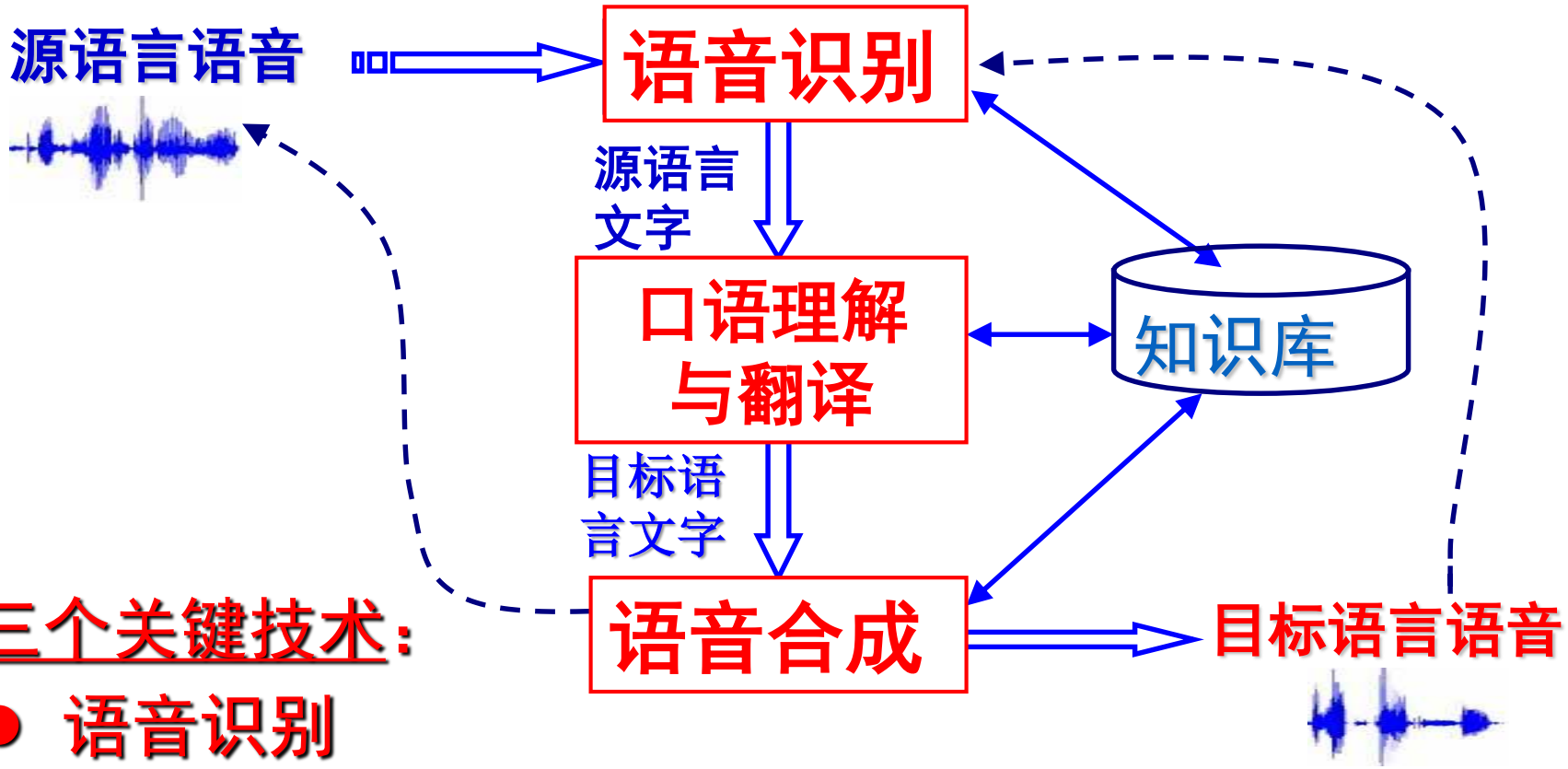
目标语言
(Target language) 2

目标语言
(Target language) 1

4. 应用举例



4. 应用举例



三个关键技术：

- 语音识别
- 口语理解和翻译
- 语音合成

4. 应用举例

- 口语翻译与书面语翻译对比

对比要素	口语翻译	书面语翻译
句长	8个词/句，1.8字/词	22.5个词/句，2.5字/词
句子结构	简单	复杂
规范性	不规范，省略、重复、修正等现象普遍存在	较规范
用词	简单，常用语大约2～3万词，少用成语	词汇量大，大约5万左右的普通词，5万以上专用词。
噪声	语音识别产生噪声	无噪声
副语言信息	重要	无

4. 应用举例

中 文 信 息 学 报
第 13 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 2

口语自动翻译系统技术评析

宗成庆　黄泰翼　徐 波
中科院自动化研究所　模式识别国家重点实验室　北京　100080

摘要　近几年来，随着信息技术的发展，口语自动翻译技术成为新的研究热点。目前国际上一些著名大学和研究机构甚至企业，都纷纷加入这一高技术的竞争行列，我国在相关技术方面也进行了卓有成效的研究。本文对目前自动口语翻译研究的技术现状进行了全面综述和分析，并对一些具体问题作了深入探讨。作者希望本文作出的分析和讨论的问题，能够对我国的自动口语翻译研究提供有益的参考。

关键词　口语翻译　语音翻译　对话处理　机器翻译　鲁棒性

一、引言

自 80 年代末期，人们开始致力于语音翻译的研究，由于该项研究不仅具有重要的科学意义，而且蕴涵着潜在的巨大社会和经济效益，因此，许多发达国家竞相投入经费开展全国性或多国性的联合攻关，美国的 CMU、日本的 ATR－ITL、德国联邦政府教育、科学、研究与技术部（BMBF）和 Siemens 等世界著名大学、研究机构和企业都是自动语音翻译研究的重要的开拓者或参与者。为了进一步推动自动语音翻译研究的更快发展，1991 年一些国际知名研究机构、大学和企业联合成立了国际口语翻译组织 C－STAR（Consortium for Speech Translation Advanced Research），到目前为止，该组织已经发展成为 C－STAR II 并拥有包括中国科学院自动化研究所、声学研究所和哈尔滨工业大学在内的二十多个会员，其研究目标是形成特定领域的口语对话翻译系统，最终实现基于自动翻译的全球自由通讯。

然而，由于自动语音翻译技术涉及语言学、声学、信号处理、模式识别和认知心理学、计算机科学等各种学科的知识和技术，同时集语音识别、机器翻译和语音合成等理论难点和技术难关于一身，使得其研究历程步履维艰，语音翻译被认为是对现代科技重要的挑战之一。

本文将对自动语音翻译研究的现状、技术特点及其存在的问题等，进行全面综述和分析，作者希望本文提出的问题及所作的分析能对自动语音翻译研究提供有益的参考。

二、口语翻译技术特点

从原理上讲，一个单向的语音翻译系统主要由语音识别、机器翻译和语音合成二个模块构成，如图 1 所示。

本课题得到国家自然科学基金资助（59835030）和国家 863 高技术项目资助（863－306－ZT03－02－2）
本文 1998 年 7 月 28 日收到

5.1　关于目前语音翻译的技术路线

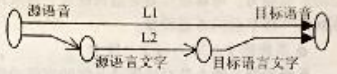


图 2　口语翻译技术路线

在目前的语音翻译系统中，都是首先将源语言的语音信号转换成文字，然后再对文字进行分析、转换、生成，最后将译文转换成语音信号输出，如图 2 中的 L2。而实际上人在进行口语翻译时，是直接从语音到语音的，尤其当翻译者对源语言比较熟悉时，根本不需要从音到字，再从字到音的转换过程，如图 2 中的 L1。

那么，在目前的语音翻译系统中，采用 L2 处理过程显然是一种不得已而为之的处理方案，而这种方案是在计算机硬、软件技术（包括内存、运算速度和知识表示方法等）受到一定局限的情况下诞生的，可以设想，当计算机的内存和运算速度足够大，语言知识包括语音知识在内的语义知识、语用知识和语境等知识得到充分表达和利用时，直接实现从语音到语音的翻译是完全有可能做到的事情。

在人们的日常对话中，情景知识是理解讲话者语音的重要因素。一个人理解对方的话语往往不只是通过对方的语言本身，而也看对方的手势、动作、表情等行为特征和讲话人的社会地位、角色以及谈话背景等抽象特征，需要时，听话人双方还要有一个问答的交互过程，在特定情况下，这些辅助信息对于帮助听众理解讲话人的内容起了决定性的作用。因此，一个面向口语的语音翻译系统应该是交互式的，系统并不一定要对讲话人的每一句话都完全“听懂”和“理解”，只要系统“理解”其中的一部分，就可以通过与用户交互来获得在字面上无法反映出来的对话情景知识，从而帮助系统对源语言句子理解和翻译。

总之，作者认为，语音翻译的技术路线不应该是单一的，必须充分体现人脑完成口语翻译过程的复杂性，并在可能实现技术的基础上，建立多种方法联合协作的语音翻译集成系统。

5.2　关于语音翻译系统的评估问题

从人的要求来看，语音翻译系统能够达到的翻译正确率越高越好，如果能达到 100% 的翻译正确率，就可以完全替代人实现全自动翻译，但实际上，这是不可能的，至少在目前没有这种可能。那么，人是否一定要有 100% 正确的翻译才能理解讲话人的意图呢？回答显然是否定的，人们在日常会话中，即使持两种不同语言的人之间的会话，也不需要清清楚楚地完全听明白对方讲的每一个字，有时只需要听清一部分内容，甚至一个词就可以完全明白对方的意思。实际上，作为一个翻译系统来说，人本身也是其中的一个组成部分（语音的接受者和理解者），而且是具有相当智能的一个组成部分，因此，从这一角度考虑，我们没有必要刻意要求系统必须达到一个目前无法实现的标准，似乎达不到某一标准就予以彻底否定，系统翻译的正确率也不应该是衡量系统好坏的唯一标准。

参照机器翻译评估的方法[16－18]，作者认为，除了系统所能实现的翻译正确率以外，评价一个口语翻译系统还应着重从如下几方面考虑：

● 系统的鲁棒性和自适应能力

系统的鲁棒性一方面指语音识别机制对环境噪声和其它各种因素的抗干扰能力，另一方面指翻译机制对来自语音识别模块的输入语句的容错能力和理解能力。自适应能力则是指系统对不同讲话人的适应性。

● 系统的扩展能力和可移植性

系统的扩展能力主要指系统在识别词汇量、能够处理的语言现象、系统知识库规模等各方

4. 应用举例

5.1 关于目前语音翻译的技术路线

在目前的语音翻译系统中，都是首先将源语言的语音信号转换成文字，然后再对文字进行分析、转换、生成，最后将译文转换成语音信号输出，如图 2 中的 L2。而实际上人在进行口语翻译时，是直接从语音到语音的，尤其当翻译者对源语音比较熟悉时，根本不需要从音到字、再从字到音的转换过程，如图 2 中的 L1。

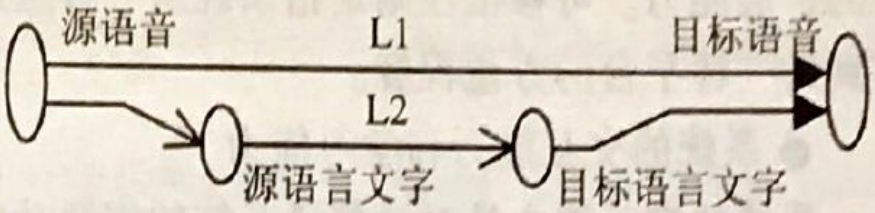


图 2　口语翻译技术路线

那么，在目前的语音翻译系统中，采用 L2 处理过程显然是一种不得已而为之的处理方案，而这种方案是在计算机硬、软件技术（包括内存、运算速度和知识表示方法等）受到一定局限的情况下诞生的，可以设想，当计算机的内存和运算速度足够大，语言知识包括语音知识在内的语义知识、语用知识和语境等知识得到充分表达和利用时，直接实现从语音到语音的翻译是完全有可能做到的事情。

宗成庆，黄泰翼，徐波，口语自动翻译系统技术评析，中文信息学报，1999年第2期，第56－65页

提 纲

1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. **技术现状**
6. 我们团队

5. 技术现状

- **NLP技术的确在很多领域得到了成功应用**
  - **汉字输入**
  - **信息查询/ 检索**
  - **人机对话（Siri/ 语音助手等）**
  - **语音识别/ 语音合成**
  - **机器翻译**
    - 百度翻译/ 腾讯翻译君/ 搜狗翻译/ GoogleTranslate/ Bing Microsoft Translator/ TRADOS ……
    - 航空等民用领域
    - 军队、安全、公安等特定领域

……

5. 技术现状

◆汉语自动分词技术现状

●以网络文本分词为例

错误类型			错误数	比例(%)			例子
集外词	命名实体	人名	31	25.83	55.0	98.33	约翰·斯坦贝克
		地名	11	9.17			米苏拉塔
		组织机构名	10	8.33			泰党
		时间和数字	14	11.67			37万兆
	专业术语		4	3.33			脱氧核糖核酸
	普通生词		48	40.00			致病原
切分歧义			2	1.67			歌名为
合计			120	100			

从互联网上随机摘取了418个句子，共含11,739个词，19,777个汉字（平均每个句长约为28个词，每个词约含1.68个汉字）。

5. 技术现状

- 以微博分词为例

类别	类别描述
事件报道	特定事件/具体事件
新闻内容	新闻消息/格式较规范
观点传播	观点词汇多/日常闲谈/观点评论
信息共享	分享的信息或者链接/为他人提供的建议
私人会话	帖子开头有“@某人”/日常闲谈
交易信息	帖子中出现金钱、比例词汇

根据对2011年微博内容的统计，大约75%的内容为个人心情和感受方面的。

5. 技术现状

补充词汇：

词典来源	词语数量
维基百科+常用在线词典	1301320
微博用语词库	10330
网络用语大全	294
网络关键词以及词频数据	500000
人民日报微博词频统计	42315
百度百科对于网络用语的解释	1051
网络用语词典	541941（经过合并筛选）
网络情感词典+传统情感词典	26207
情感词典	26207
词语总数：1,753,925（经过合并筛选）	

5. 技术现状

分词性能：

分词方法	准确率(%)	召回率(%)	F1值(%)
Stanford	80.40	76.52	78.41
Urheen	80.46	77.43	78.92
ICTCLAS(+微博处理)	82.62	83.52	83.07
CWS	80.12	73.24	76.52
CWS(+词典+符号处理)	90.52	90.73	90.62

CWS: Chinese word segmentation based on ME model

5. 技术现状

- **以半文言文文本分词为例**

  李时珍（约1518～1593），字东璧，晚号濒湖山人，蕲州（今湖北蕲春）人。世业医，父言闻，有医名。幼习儒，三次应乡试不中。自嘉靖三十一年（1552年）至万历六年（1578年），历时二十七载，三易其稿，著成《本草纲目》五十二卷，初刊于金陵。

  分词准确率为：57.3%～94.8%

- **问题归纳：**
  - 生词识别和切分是汉语自动分词技术面临的最大问题
  - 跨领域和非规范是导致生词大量出现的主要原因
  - 研究半监督学习、迁移学习等方法，解决领域的自适应问题，提高系统的鲁棒性和准确率，尽量减少系统对标注样本的依赖性，是未来汉语自动分词技术研究的主要方向

5. 技术现状

◆机器翻译译文质量得到明显提高

原文：Beijing made a third solemn representation to Manila and warned that it is hard to be optimistic about a territorial impasse over an island. Authorities say they have prepared for any escalation of the situation by Manila.

(http://www.chinadaily.com.cn/, 8 May 2012)

Results given by Google Translator：

5. 技术现状

2012.5.8 …

- 北京做第三严正交涉到马尼拉，并警告说，这是很难约为领土僵局在一个小岛乐观。当局说，他们已经准备的情况下由马尼拉的任何升级。(2014.4.24)
- 北京做第三严正交涉到马尼拉，并警告说这是很难约领土僵局的一个岛屿乐观。当局说，他们已经准备了马尼拉的情况有任何升级。(2015.4.28)
- 北京由第三严正交涉到马尼拉，并警告说这是很难约了一个岛领土僵局持乐观态度。当局说，他们已经为马尼拉局势的升级准备。(2016.5.1)
- 北京对马尼拉进行了第三次庄严的代表，并警告说，对岛上的领土僵局很难看好。当局表示，他们为马尼拉的情况升级做好了准备。(2017.5.15)
- 北京向马尼拉提出了第三次严正交涉，并警告说很难对一个岛屿的领土僵局持乐观态度。当局表示，他们已经为马尼拉局势升级做好了准备。(2018.5.20)
- 北京向马尼拉提出了第三次庄严的代表，并警告说，对一个岛屿的领土僵局很难保持乐观。有关当局表示，他们已经为马尼拉的局势升级做好了准备。 (2019.5.6)

5. 技术现状

●仍存在太多的问题

文本翻译实例

① 新华社巴姆1月2日电(记者蒋晓峰廖振云)正在伊朗巴姆地区参加地震救灾工作的美国国际开发署人道援助项目主管史蒂夫·卡特林, 2日对新华社记者说,美国救援队到伊朗参加救援活动不带任何政治色彩 。

美伊两国自1980年断交以来,双方一直没有恢复正常关系。

Xinhua News Agency, Bam, January 2 (Reporter Jiang Xiaofeng Liao Zhenyun) Steve Catlin, director of the US Agency for International Development Humanitarian Assistance Project, who is participating in the earthquake relief work in the Bam area of Iran, told the Xinhua News Agency on the 2nd that the US rescue team There is no political color in going to Iran to participate in rescue activities.

Since the United States and Iraq broke diplomatic relations in 1980, the two sides have not resumed normal relations.

- Google Translator, 2018.12.20

难以准确地处理篇章范围内的指代问题。

5. 技术现状

② 阿Q没有理由怕老婆。 ⇨ Ah Q has no reason to be afraid of his wife.

阿Q没有理由地怕老婆。⇨ Ah Q has no reason to be afraid of his wife.

无法准确区分由于细微文字或句法差异造成的句子语义反转。

③ 新加坡经检会：2004年经济可望全面复苏。

Singapore Economic Inspection Committee: The economy is expected to fully recover in 2004.

Singapore's Economic Review Committee: The economy is expected to fully recover by 2004.

无法从译员的译后编辑过程中自动学习翻译知识。

5. 技术现状

④《习近平谈“一带一路”》收入的42篇重要文稿，深刻阐述了“一带一路”建设的重大意义、指导原则、丰富内涵、目标路径等重大问题，对于推动“一带一路”国际合作不断取得新进展，造福沿线各国人民，推动构建人类命运共同体，具有十分重要的指导意义。

"Xi Jinping talks about the" One Belt, One Road " revenue of 42 important manuscripts, profoundly elaborated the " Belt and Road "construction of the significance of the major issues, guiding principles, rich connotations, the target path and other major problems, to promote the " Belt and Road " international cooperation continues to make new progress for the benefit of the people along the route, and promote the construction Has a very important guiding significance. **(BMT, Mar. 24, 2019)**

在复杂长句翻译中大量地出现漏翻、错翻和词序颠倒等问题。

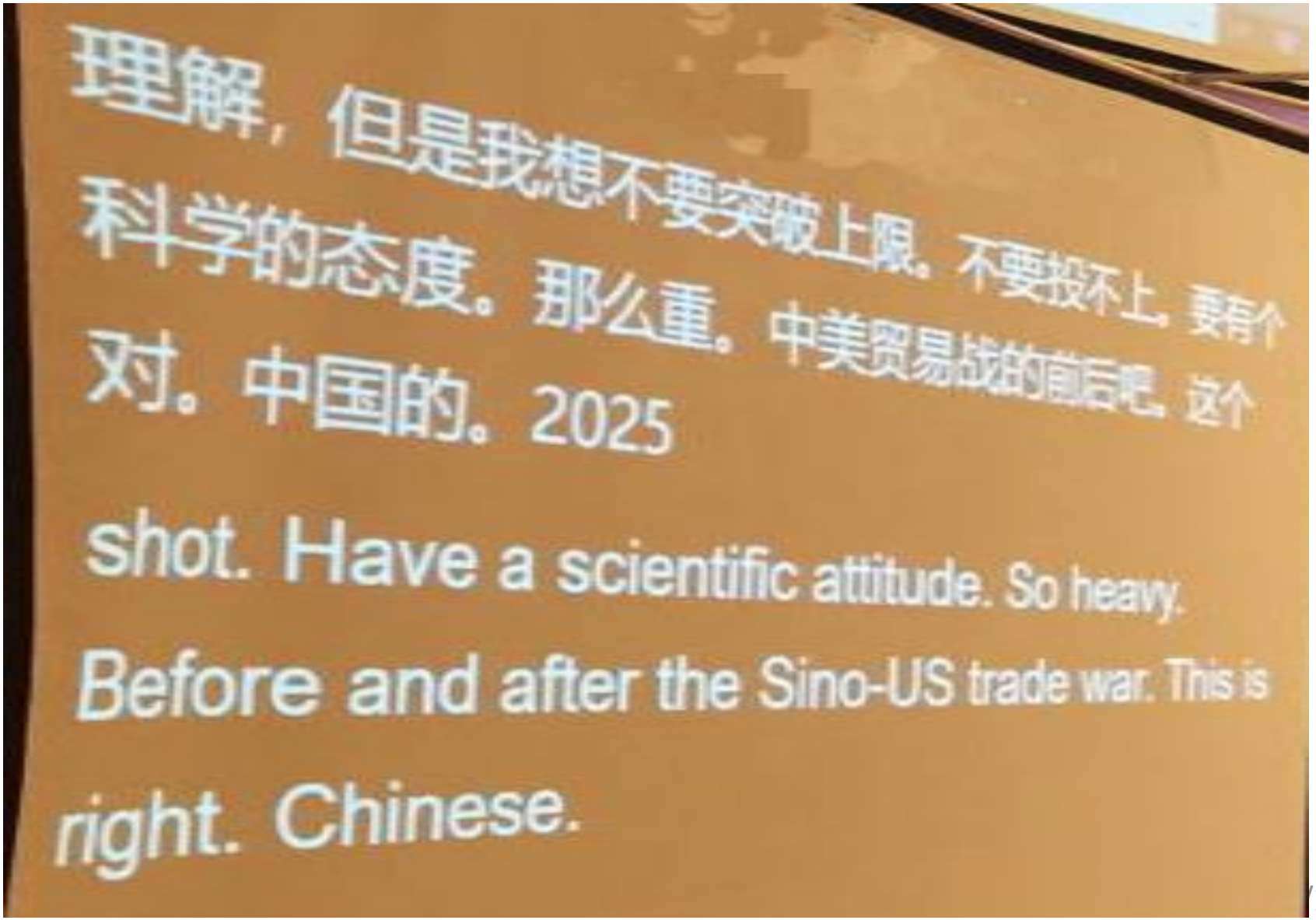
5. 技术现状

⑤ 面对原文中的噪声常常束手无策

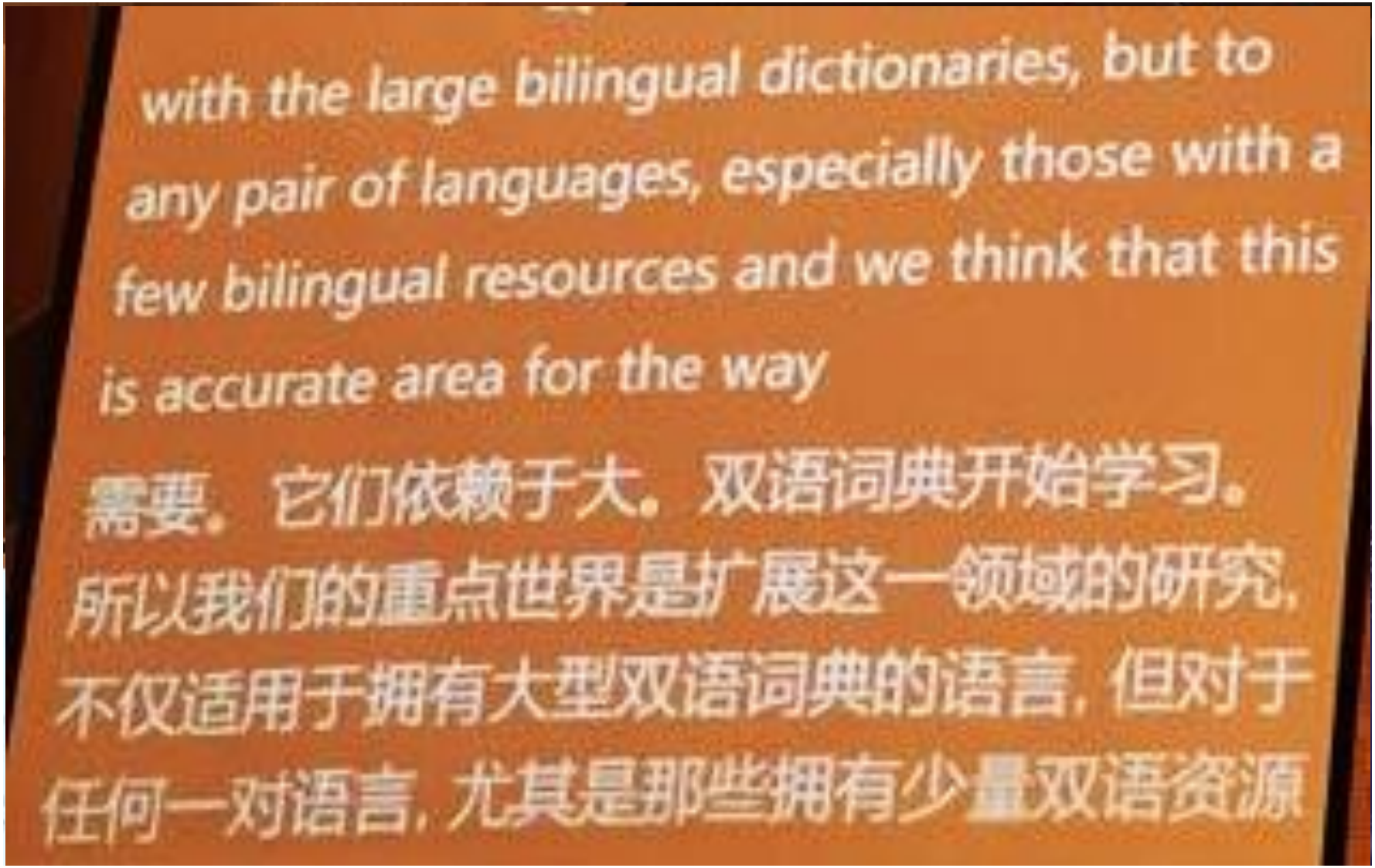
印尼语、马来语、菲律宾语三种语言的主流媒体中新闻文本分别有20%、14%和10%以上的词汇存在拼写错误。

如印尼语中表示“失败”的词“gagal”，常被“ga2l”替代；为了强调语气，有时会重复输入字符，“cemungudh”常被写为“cemunguuuuudh”。

5. 技术现状


理解，但是我想不要突破上限。不要投不上。要有个
科学的态度。那么重。中美贸易战的前后吧。这个
对。中国的。2025
shot. Have a scientific attitude. So heavy.
Before and after the Sino-US trade war. This is
right. Chinese.

5. 技术现状


with the large bilingual dictionaries, but to
any pair of languages, especially those with a
few bilingual resources and we think that this
is accurate area for the way
需要。它们依赖于大。双语词典开始学习。
所以我们的重点世界是扩展这一领域的研究,
不仅适用于拥有大型双语词典的语言,但对于
任何一对语言,尤其是那些拥有少量双语资源

5. 技术现状

◆性能现状

- 对于资源较为充分的语言对（如英汉、日汉等），在说话场景不是非常复杂、口音基本标准、语速基本正常、使用词汇和句型不是非常生僻的情况下，日常口语翻译的性能基本可满足正常交流的需要；专业领域的文本翻译在训练语料较充分时译文准确率可达到80%以上；新闻领域的翻译准确率跨度较大，总体而言，文本翻译的准确率不足75%。
- 对于译文质量要求较高的翻译任务，如领导人的讲话稿或著作、文学名著、严肃场景下的演讲和对话（包括领导人讲话、外交部答记者问、有较严重口音的学术讲座等），机器翻译系统都难以胜任。在可预见的未来看不到机器翻译系统会替代人工翻译的可能性。
- 对于资源稀少的小语种与汉语之间的翻译，机器翻译系统只能以快速获取信息为目的帮助人们大致了解原文的主题和内容。

5. 技术现状

◆问题归纳

- 先验知识未能得到充分利用，歧义结构和多义词汇不能得到准确地翻译
  - 词典、规则很难融入神经机器翻译模型
  - 仅从双语句对中难以获取、表示和利用背景知识及常识

    例如：transformers: ①改革者；②变形金刚；③变压器。
- 生词(成语、命名实体等)仍然是当前机器翻译面临的主要问题

  例如：天行健，君子以自强不息；地势坤，君子以厚德载物。

  事者，生于虑，成于务，失于傲。

  四十载惊涛拍岸，九万里风鹏正举。

  (习近平主席在纪念改革开放40周年大会上的讲话)
- 语义（话语意图）理解是最终攻克机器翻译难关的瓶颈

5. 技术现状

16/content_32646723.htm?from=groupmessage&isappinstalled=0

光明日报 时政 国际 时评 理论 文化 科技 教育

首页> 光明日报 > 正文

机器翻译的梦想与现实

来源：光明网-《光明日报》 2019-03-16 04:54

作者：宗成庆（中国科学院自动化研究所研究员）

主持人语：

近年来，机器翻译等语言智能技术迅猛发展，并开始在舆论中对人工外语服务形成冲击。一时间，语言智能技术落地、机器翻译替代人工、外语教育存废、外语教学改革等，成为社会热议的话题。本期聚焦“语言智能与外语教育协同发展”，邀请多领域专家学者共同探究，外语教育如何将语言智能的挑战作为转型的契机更好发展，怎样理性认识语言智能技术的发展未来，全力适应人机共存的语言生活形态，努力过好智能时代的外语生活，从而推动语言智能和外语教育协同发展，共同进步，共同助力人类命运共同体的形成。

——李宇明

光明网：http://share.gmw.cn/news/2019-03/16/content_32646723.htm

5. 技术现状

◆ 远未做到语言的深度理解

高山从警20多年来，历尽千辛万苦，立下无数战功，曾被誉为孤胆英雄。然而，谁也未曾想到，就是这样一位曾让毒贩闻风丧胆的铁骨英雄竟然为了区区小利而铤而走险，悔恨之下昨晚在家开枪自毙。

- 高山是什么警察？
- 高山死了没有？
- 高山为什么自杀？

5. 技术现状

◆ 远未做到语言的深度理解

Name　　Time

高山/ 从警/ 20多年/ 来/ ，/ 历尽/ 千辛万苦/ ，/ 立下/ 无数/ 战功/ ，/ 曾/ 被/ 誉为/ 孤胆英雄/ 。/ ‖然而/ ，/ 谁/ 也/ 未/ 曾/ 想到/ ，/ 就是/ 这样/ 一位/ 曾/ 让/ 毒贩/ 闻风丧胆/ 的/ 铁骨/ 英雄/ 竟然/ 为了/ 区区/ 小利/ 而/ 铤而走险/ ，/ 悔恨/ 之下/ 昨晚/ 在/ 家/ 开枪/ 自毙/ 。

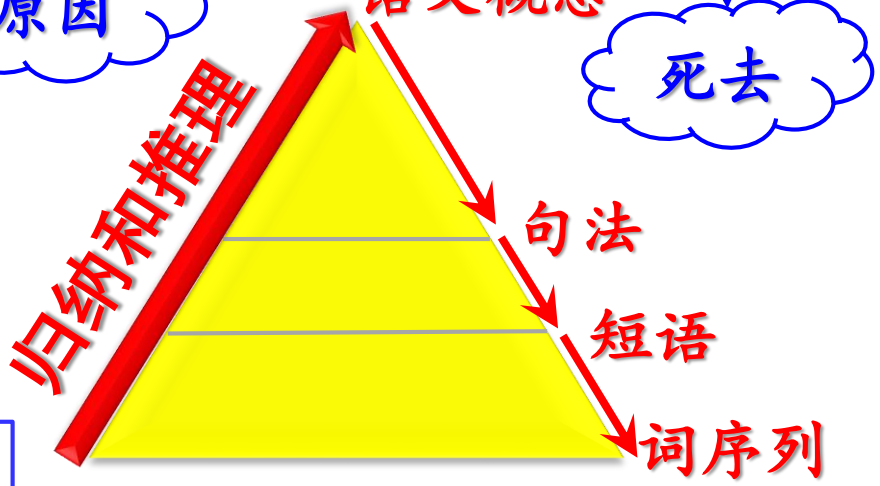
缉毒警察　　原因　　语义概念　　死去

① 分词（96%）

② 命名实体识别（90%）

③ 实体关系抽取（85%）

④ 语义角色标注（70-82%）

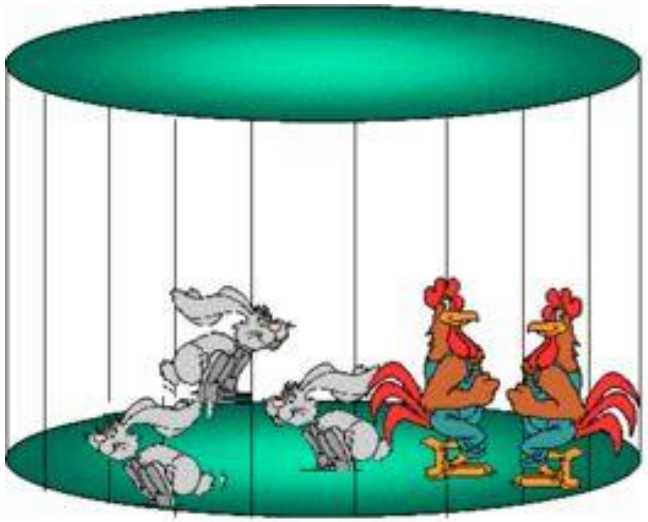


目前端到端的方法缺乏推理能力

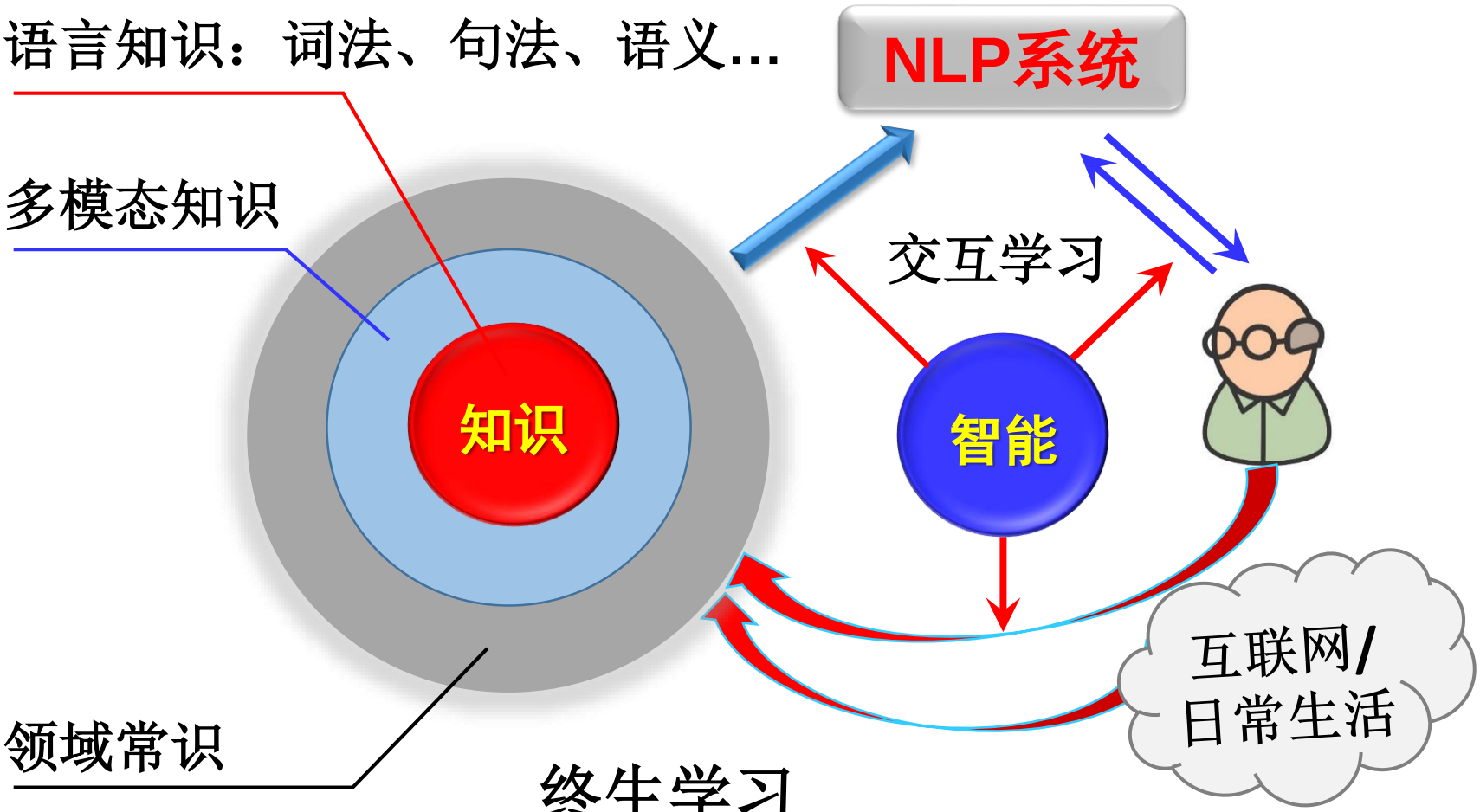
5. 技术现状

一群鸡和兔子，放在同一个笼子里，上面有35个头，下面有94只脚，问有多少只鸡、多少只兔？

- 焦点词确定
- 常识获取
- 交互学习



5. 技术现状



终生学习
Lifelong learning

5. 技术现状

- 包括深度学习在内的统计学习方法毕竟是一种“赌博”思维，只能处理大概率事件，且其性能表现严重依赖于训练样本，难以做到“举一反三”，基于大规模训练样本建立的系统性能往往不如一个三岁的小孩，尤其缺乏合理的解释。
- 多源信息的融合分析与理解成为必然趋势
  - 多语言文本理解
  - 图像、视频内容分析与理解
  - 语音识别与说话人识别
  - 多源信息融合方法与问题求解
  - 人机交互

提 纲

1. 学科产生与发展
2. 技术挑战
3. 基本方法
4. 应用举例
5. 技术现状
6. 我们团队

6. 我们团队

◆ 主要研究方向

- 自然语言处理基础 (*Fundamental research on NLP*)
  - 汉语自动分词 (*Chinese word segmentation*)
  - 自动句法分析 (*Parser*)
  - 语义角色标注 (*Semantic role labeling*)
  - 命名实体识别 (*Named entity recognition*) ……
- 机器翻译 (*Text and spoken language translation*)
- 知识工程/ 知识图谱
- 信息抽取和问答系统 (*IE/Q&A*)
- 文本分类/情感分类 (*text/sentiment classification*)
- 多模态信息融合的自然语言处理 (*Multimodal NLP*)
- 类脑启发的自然语言理解 (*Brian inspired NLU*)

6. 我们团队

◆多语言机器翻译及分析系统

中科院自动化所模式识别国家重点实验室自然语言处理团队研发的“**中科凡语**”多语言机器翻译及分析系统已实现18种语言、20种文字到汉语的自动翻译和分析，成功应用于多个特定领域。

6. 我们团队

知识服务应用

医疗

司法

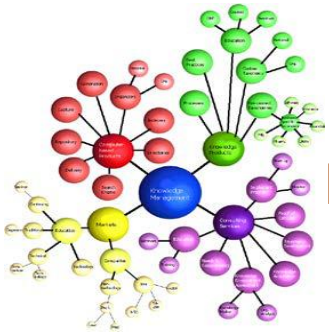
电商

金融

教育

大规模开放域知识获取与知识图谱构建

- 实体识别
- 属性抽取
- 实体消歧
- 事件抽取
- 关系抽取
- 事件关系

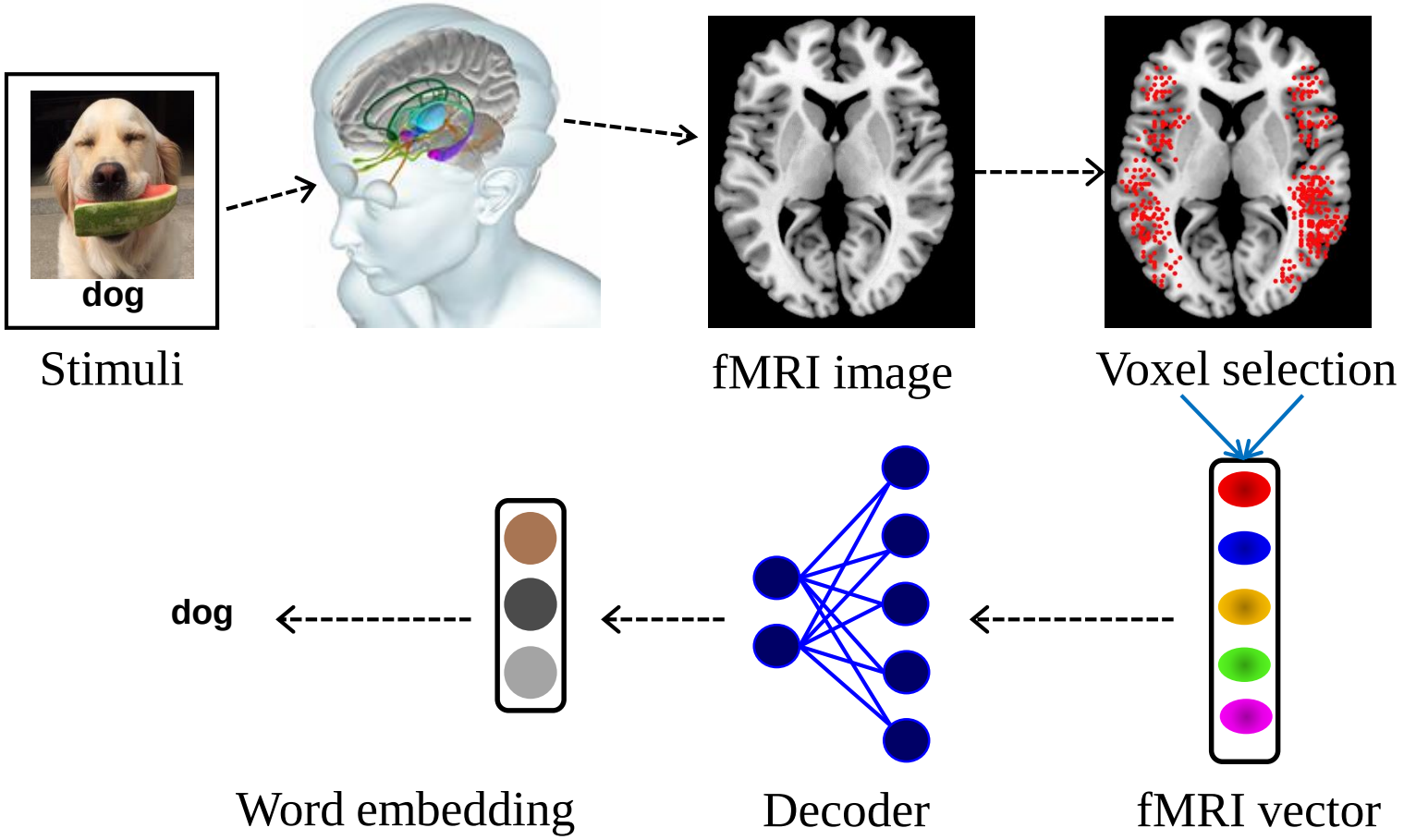


知识图谱

面向大规模、多源异构知识图谱的深度问答

- 问句分类
- 语义解析
- 知识匹配
- 知识推理
- 阅读理解
- 信息验证

自然语言处理基础技术

6. 我们团队

◆ 受脑启发的自然语言语义表示、编码及解码方法