


自然语言处理方法与应用

宗成庆

中国科学院自动化研究所
模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

内容提要

-  1. 问题的提出
- 2. 自然语言处理方法
- 3. 应用举例
- 4. 技术现状
- 5. 结束语

1. 问题的提出

◆ 什么是自然语言？

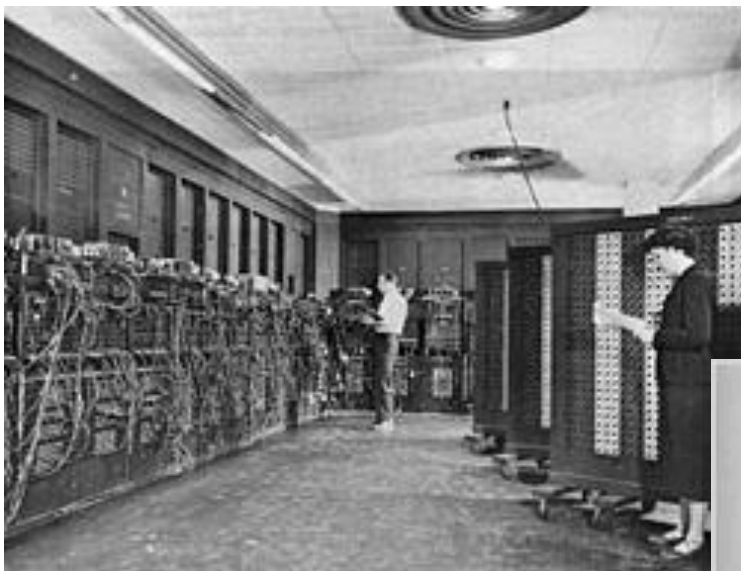
- 自然语言是人类社会发展过程中自然产生的语言，是最能体现人类智慧和文明的产物



- 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. 问题的提出

◆ 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. 问题的提出



[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



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

1. 问题的提出

- 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。

1. 问题的提出



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。

1. 问题的提出

- 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。

1. 问题的提出

- 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

1. 问题的提出

◆ 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）

1. 问题的提出

◆ 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）

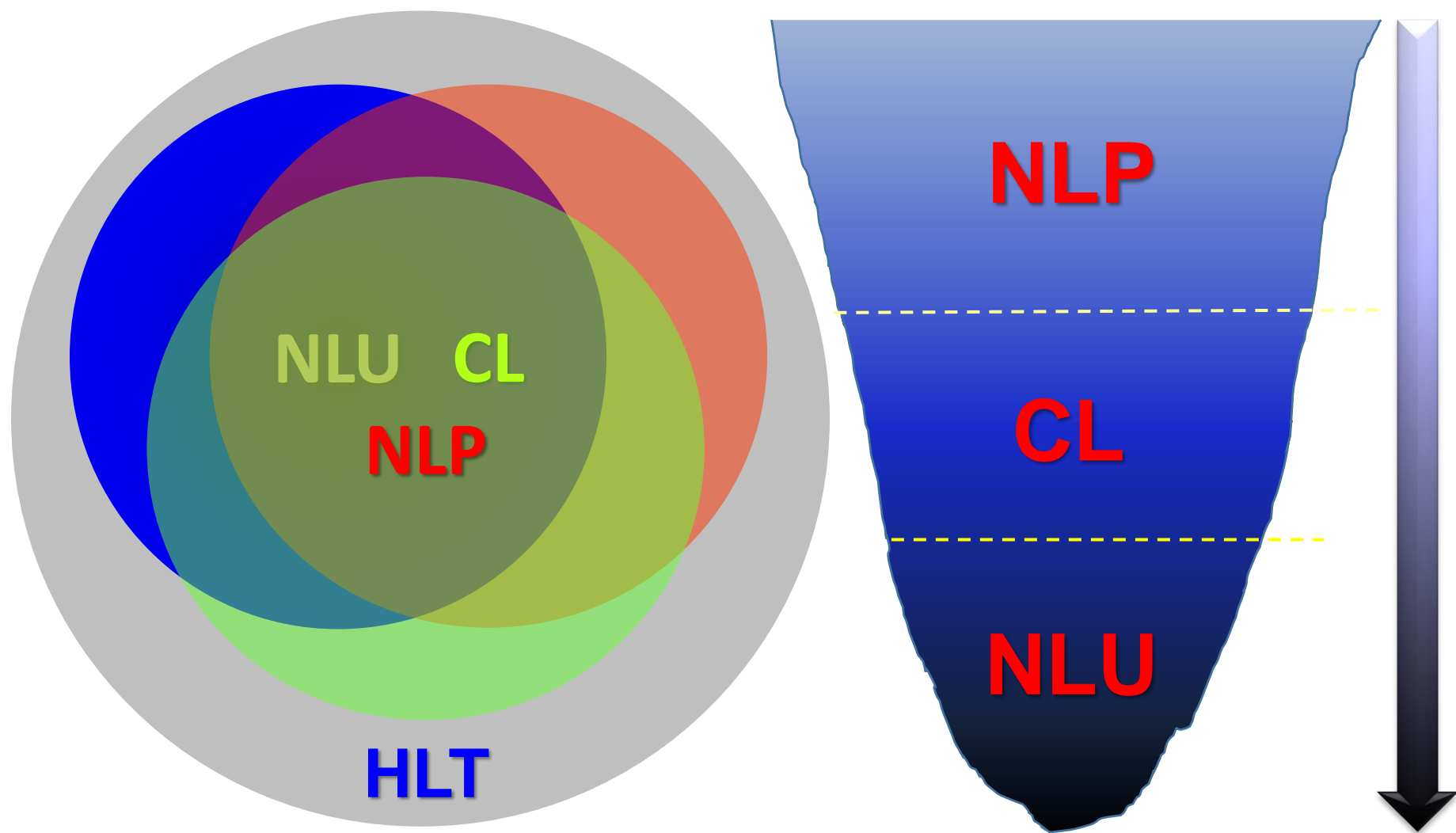
1. 问题的提出

◆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）

1. 问题的提出

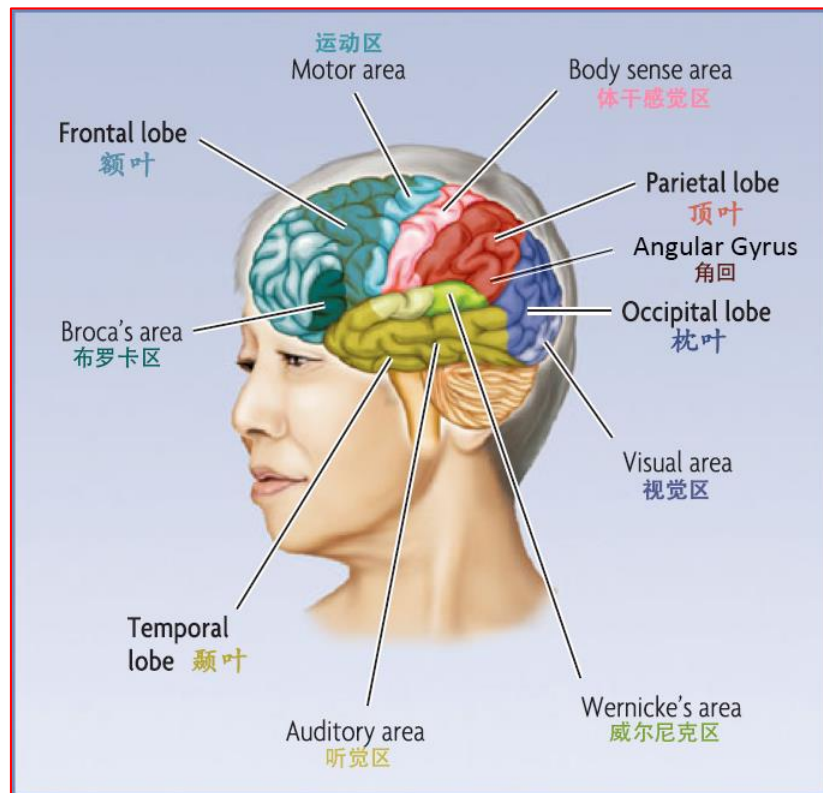


统称：人类语言技术 (human language technique, HLT)

1. 问题的提出

◆ 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
- 推动相关学科发展



1. 问题的提出

◆ 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. 问题的提出

360安全浏览器 3.18 正式版

http://news.sina.com.cn/s/2012-03-08/233624083315.shtml

文件(E) 查看(V) 收藏(B) 账户(U) 工具(T) 帮助(H)

后退 前进 停止 刷新 主页 恢复 收藏 历史 无痕

网络收藏夹 添加收藏 常用 链接 MSN.com 电台指南 Microsoft 网站 Windows Live 服务 Coling 2010 Home ACL Anthology -A...

文本

扫描文档

视频

图片

Flash

中新网
ChinaNews.com

3月1日事发现场。左2为一再喊冤的广西“许云鹤”张都。 林增崇 摄

中新网南宁3月8日电(林增崇 孙洁)被炒得沸沸扬扬的广西版“许云鹤”事件目前水落石出。一名自称“学雷锋”扶倒地老人而被冤枉成肇事者的广西玉林男子,在交警部门展示其撞车视频后,终于不再四处喊冤,承认自己就是肇事者。

都,来自广西博白,我是一个农村进城的务工人员。3月1日,我在博白县发生这样一件事。做好事不留名,结果却让我背上了“肇事者”的罪名。他受伤又关我何事?他死了又关我何事?他发生了天津的“许云鹤案”吗?我无助,所以我把我遭遇的事情公之于众,我要让更多人知道,把事情说清楚。还我一个清白,还社会一个公道。再寒冷,我也要温暖。伟大领袖毛主席号召向雷锋同志学习49周年,人们行动起来,不要让“彭宇案”成为“许云鹤案”。人们做好事而受到冤枉,他们太不应该付出这样的代价了。

3月2日,天涯社区、广西红豆社区等论坛出现了《“许云鹤案”再现广西??》的帖子,作者自称“张都”,是广西玉林市博白县人,他以“做好事被交警和当事人家属冤枉为肇事者”为由发帖,寻求网上舆论支持:

“3月1日上午11点40分左右,我开着自家的面包车从玉林大北路西城停车场出来,在大门口我看见一个老人站在马路中间的隔离栏边

热门博客

“红泥人”的奇异婚俗 地狱般的印度陨石矿场

买个秦始皇的村官要花多少钱(图) Qing
毛泽东何时下决心打倒刘少奇(图)
彭德怀如何交代百团大战“罪行”
林彪要求对叶群是处女的两个证据
男人有外遇跟老婆好不好没有关系
微博发现卖萌考驾照 发微博曝光迟到者
太阳风暴袭击地球 GPS与飞机航行受影响
The new iPad发布 多项改进你会买吗
湖南卫视《锋尚之王》票选微公益

智投导购

探秘领导一个月
搞定英语
出国不用翻译

查看详情

教育 教育 教育 教育

解酒保肝一不醉有秘诀 赚时尚男女财富的绝招

完成

开始 收件箱 - Outlook

87.8%的网络内容为非结构化文本。

1. 问题的提出

360安全浏览器 3.18 正式版

http://news.sina.com.cn/s/2012-03-08/233624083315.shtml

文件(E) 查看(V) 收藏(B) 账户(U) 工具(T) 帮助(H)

后退 前进 停止 刷新 主页 恢复 收藏 历史 无痕

网络收藏夹 添加收藏 常用 链接 MSN.com 电台指南 Microsoft 网站 Windows Live 服务 Coling 2010 Home ACL Anthology -A...

热门博客

- “红泥人”的奇异婚俗
- 地狱般的印度陨石矿场
- 买个秦始皇的村官要花多少钱(图) Qing
- 毛泽东何时下决心打倒刘少奇(图)
- 彭德怀如何变代百团大战“罪行”
- 林彪要求对叶群是处女的两个证据
- 男人有外遇跟老练不好没有关系
- 微博发现卖萌考勤机 发微博曝光迟到者
- 太阳风暴袭击地球 GPS与飞机航行受影响

3月1日事发现场。左2为一再喊冤的广西“许云鹤”张都。 林增崇 摄

中新网 Chinanews.com

3月1日事发现场。左2为一再喊冤的广西“许云鹤”张都。 林增崇 摄

中新网南宁3月8日电(林增崇 孙洁) 被炒得沸沸扬扬的广西版“许云鹤”案，终于不再四处喊冤，承认自己就是肇事者。

都，来自广西博白，我是一个人，这样的事发生在我身上，关我何事？他受伤又发生天津的“许云鹤案”，助，所以我把我的事情说清楚。还我一个再寒冷。

伟大领袖毛主席号召向雷锋同志学习，人们行动起来，不要让“彭宇案”让人们做好事而要付出他们本不该付出的代价。

一个老人站在

完成

开始 收件箱 - Outlook

87.8%的网络内容为非结构化文本。

文本

图片

Flash

扫描文档

机器翻译

自动摘要

观点挖掘

信息抽取

自动问答

情感/情绪分析

人机对话

视频

1. 问题的提出



1. 问题的提出



1. 问题的提出

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. 问题的提出

◆ 问题与挑战

- 大量的未知现象

如：高山, 虎蝇, 埃博拉

- 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.

- 普遍存在的隐喻表达

在微信圈里潜水

- 跨语言语义概念不对等

馒头：steamed bread

句子：We do chicken right.



(1) 张三今天中午吃馒头。

(2) 李四今天中午吃食堂。

(3) 王五今天中午吃大碗。

大量存在的不确定性和语义概念表示与计算的复杂性

内容提要

1. 问题的提出

 2. 自然语言处理方法

3. 应用举例

4. 技术现状

5. 结束语

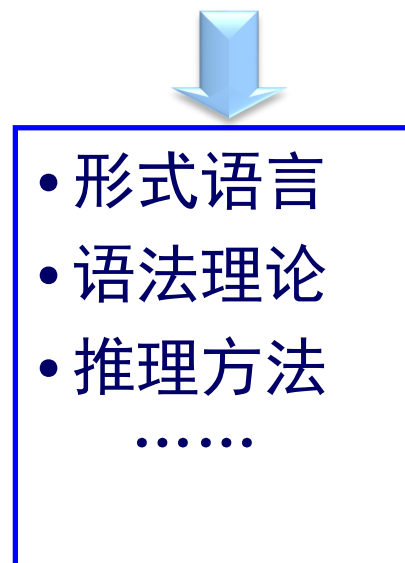
2. 自然语言处理方法

2.1 方法概述

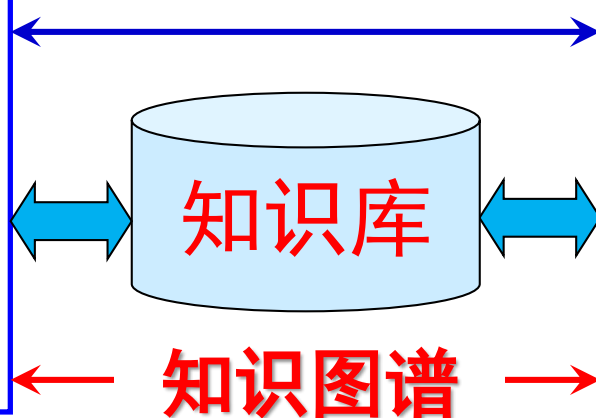
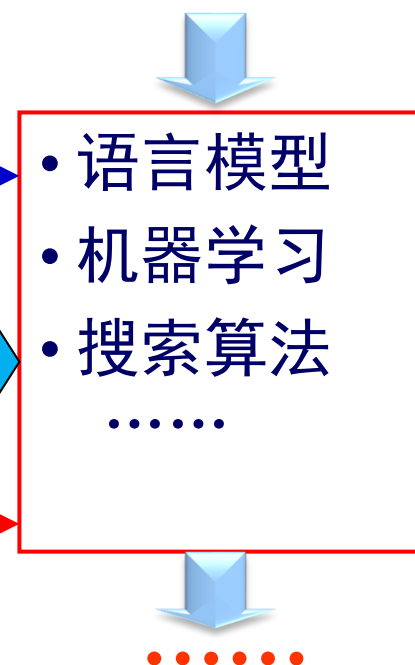
- 理性主义方法：1957~1980S
 - 词法分析，句法方法，语义分析
 - 词典、规则—**基于规则的方法**
- 经验主义方法：~1950S，1980S~
 - 大规模训练样本
 - 数据驱动的统计模型—**基于统计的方法**

2. 自然语言处理方法

基于规则的方法



基于统计的方法



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

2. 自然语言处理方法

2.2 传统的统计方法 (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

2. 自然语言处理方法

◆ 统计学习方法

- ◇ 数据驱动
- ◇ 对数据进行预测与分析
- ◇ 以方法为中心，构建模型

◆ 统计学习类型：

- 监督学习(supervised learning)
- 非监督学习(unsupervised learning)
- 半监督学习(semi-supervised learning)
- 强化学习(reinforcement learning)

2. 自然语言处理方法

● 监督学习(supervised learning)

- 给定有限的、人工标注好的大量数据，假设这些数据是独立同分布产生的(训练集, training data)
- 假设要学习的模型属于某个函数的集合，即假设空间(hypothesis space)
- 应用某（些）个评价准则(evaluation criterion)，从假设空间中选取最优的模型，使其对已知的训练数据和未知的测试数据(test data)在给定的评价准则下有最优的预测

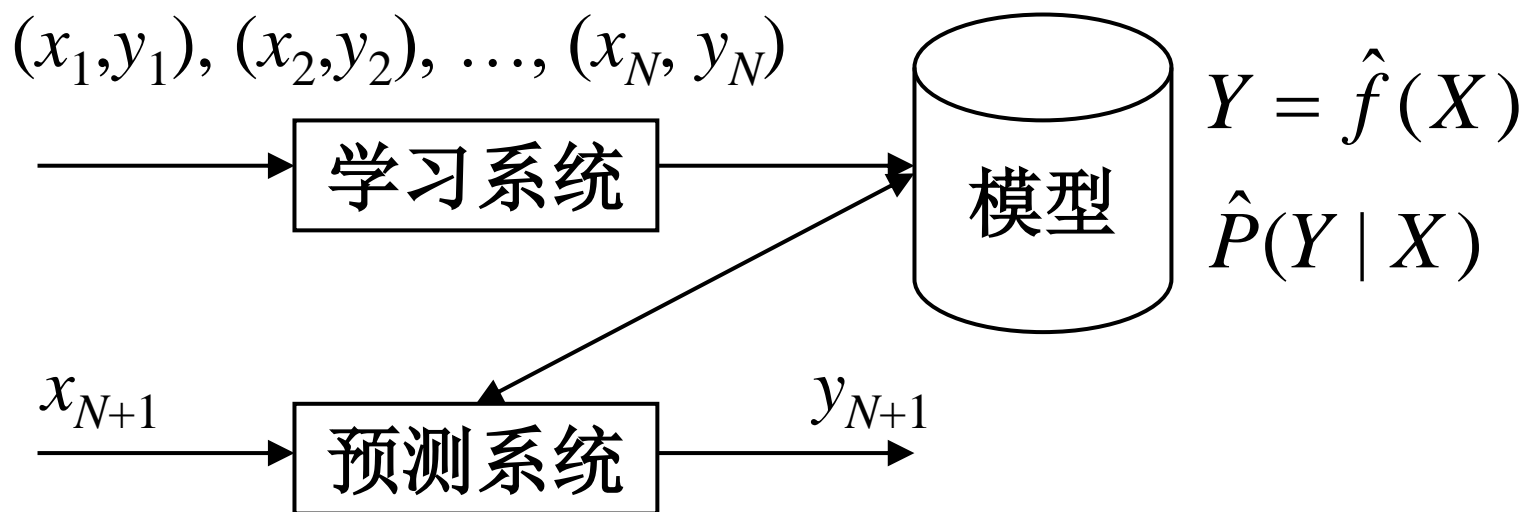
2. 自然语言处理方法

一般步骤:

- ① 获得一个有限的训练数据集合
- ② 确定包含所有可能的模型的假设空间，即学习模型的集合
- ③ 确定模型选择的准则，即学习的策略
- ④ 通过学习方法选择最优模型
- ⑤ 利用学习到的最优模型对新数据进行预测或分析

2. 自然语言处理方法

问题的形式化:



给定一个训练数据集: $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中, (x_i, y_i) , $i=1, 2, \dots, N$, 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。

2. 自然语言处理方法

◆ 模型的类别

- 生成式方法 (generative model)
- 区分式方法/判别式方法 (Discriminative Model)

2. 自然语言处理方法

● 生成式方法 (generative model)

假设 o 是观察值, q 是模型, 生成式方法对 $p(o|q)$ 进行建模。其基本思路是: 首先建立样本的概率密度模型, 然后利用模型进行推理预测。要求已知样本无穷多或者尽可能地多, 方法建立在统计学和 Bayes 理论的基础之上。

代表性方法:

- n 元语法模型(n -gram) / 语言模型(language model, LM)
- 隐马尔可夫模型(Hidden Markov Model, HMM)

2. 自然语言处理方法

➤ 语言模型(n -gram)的提出

如何计算一段文字(句子)的概率?

阳春三月春意盎然，少先队员脸上荡漾着喜悦的笑容，鲜艳的红领巾在他们的胸前迎风飘扬。

- 以一段文字(句子)为单位统计相对频率?
- 根据句子构成单位的概率计算联合概率?

$$p(w_1) \times p(w_2) \times \dots \times p(w_n)$$

2. 自然语言处理方法

- 更合理的计算方法

对于语句 $s = w_1 w_2 \dots w_m$ (含 m 个“词”), 其概率为:

$$p(s) = p(w_1) \times p(w_2/w_1) \times p(w_3/w_1 w_2) \times \dots \times p(w_m/w_1 \dots w_{m-1})$$

$$= \prod_{i=1}^m p(w_i | w_1 \dots w_{i-1})$$

当 $i=1$ 时, $p(w_1 | \langle BOS \rangle) = p(w_1)$ 。

一般地,

$$p(s) = \prod_{i=1}^{m+1} p(w_i | w_{i-n+1}^{i-1}) \quad (n \geq 2)$$

当 $n=1$ 时, 单个“词”称为一元文法(uni-gram); $n=2$ 时, 两个并列的同现“词”称2元文法(bi-gram); $n=3$ 时, 3个并列的同现“词”称3元文法(tri-gram)。依此类推。需解决数据平滑问题。

2. 自然语言处理方法

● 区分式/判别式方法 (Discriminative Model)

假设 o 是观察值, q 是模型, 区分式方法对 $p(q|o)$ 进行建模。其基本思路是: 在有限样本条件下建立判别函数, 不考虑样本的产生模型, 直接研究预测模型, 寻找不同类别之间的最优分类面, 反映的是不同类别数据之间的差异性。

代表性方法:

➤ 各种分类器模型

2. 自然语言处理方法

2.3 常用的统计模型和开源工具

◆统计模型

- 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(naïve 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

2. 自然语言处理方法

◆ 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/>

2. 自然语言处理方法

● 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/>

2. 自然语言处理方法

- 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):

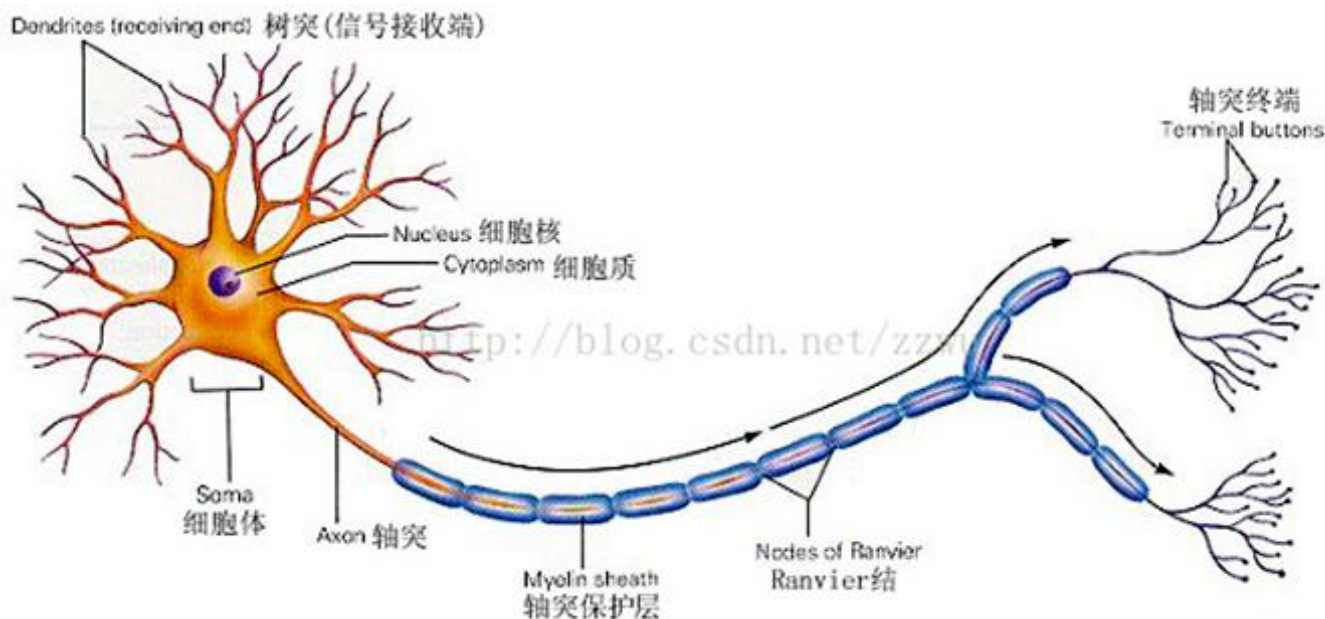
- <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

2. 自然语言处理方法

2.4 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

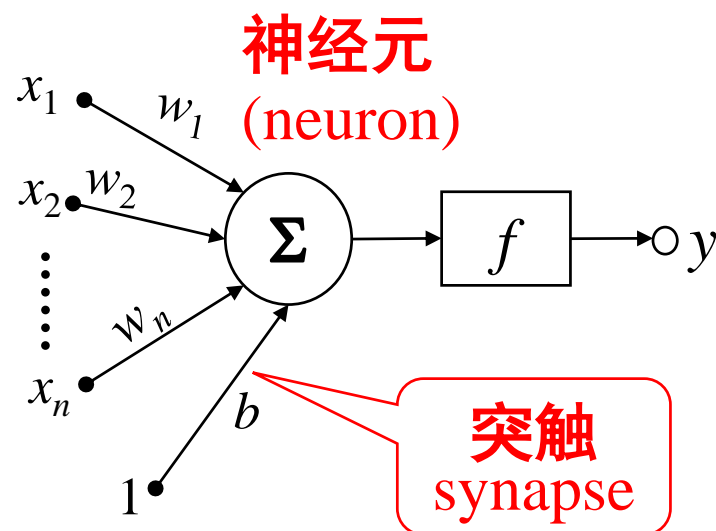


2. 自然语言处理方法

● 神经元数学描述

- $x_1 \sim x_n$ 为输入向量的各分量
- $w_1 \sim w_n$ 为权值
- b 为偏置
- f 为传递函数，通常为非线性函数
- y 为输出
- 数学表示: $y = f(\vec{W} \bullet \vec{X}' + b)$

\vec{W} 为权值向量; \vec{X} 为输入向量, \vec{X}' 为 \vec{X} 的转置。



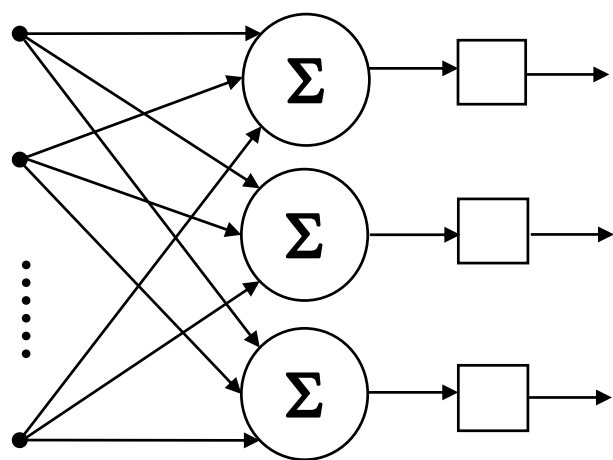
2. 自然语言处理方法

● 神经网络 (neural network)

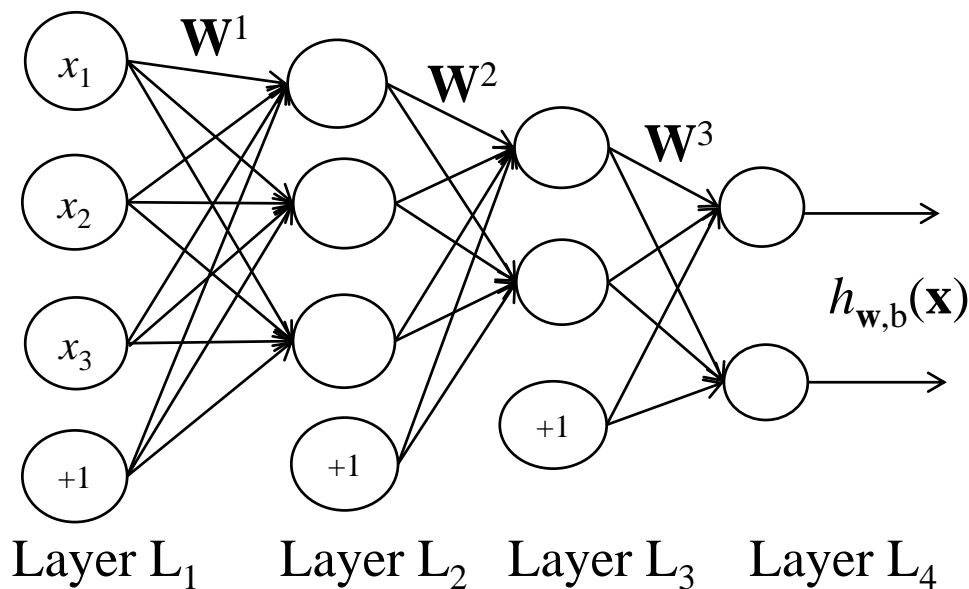
1982, Hopfield 提出神经网络模型;

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

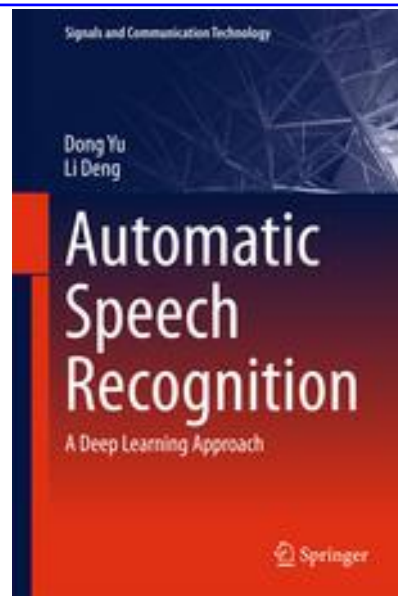
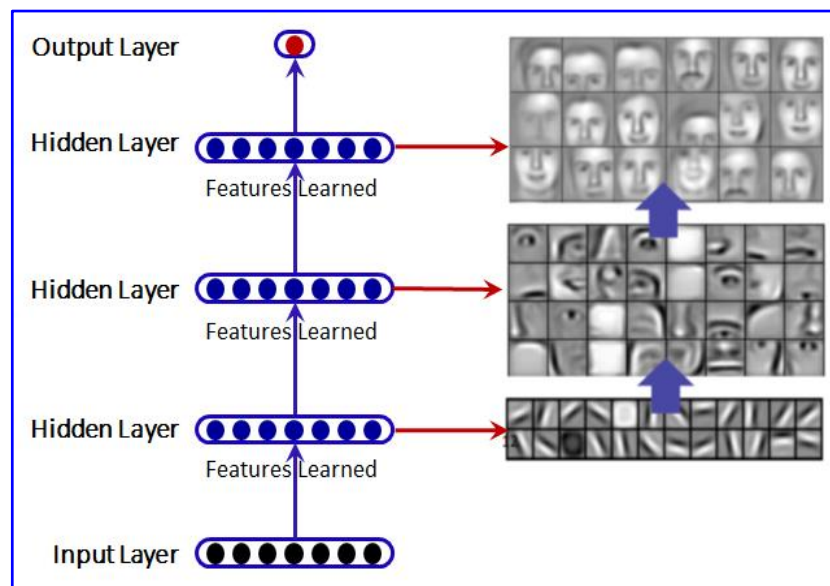


单层网络



2. 自然语言处理方法

- 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 Boltzman 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



2. 自然语言处理方法

◆ 神经语言模型

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



$$\begin{aligned} & p(w_i | w_{i-1} \cdots w_{i-n+1}) \\ &= \frac{\text{count}(w_i w_{i-1} \cdots w_{i-n+1})}{\text{count}(w_{i-1} \cdots w_{i-n+1})} \end{aligned}$$

2. 自然语言处理方法

s: 这个人说话很 风趣，大家觉得他很 幽默。

$$p(\text{风趣}|\text{很}) = \frac{\text{count}(\text{很 风趣})}{\text{count}(\text{很})}$$

问题:

- ①可能出现数据稀疏现象： n 元组“很 幽默”可能未出现过
- ②忽略了词语之间的语义相似性：“幽默”与“风趣”具有很大的语义相似性，但在表层统计意义上却无法共享信息

$$p(\text{风趣}|\text{很}) \approx p(\text{幽默}|\text{很}) ?$$

2. 自然语言处理方法

➤ 曾经的方法：抽象符号（字符串）

这个人说话很 **风趣**，大家觉得他很 **幽默**。

w_0 =这个 w_1 =人 w_2 =说话 w_3 =很 w_4 =风趣 w_5 =，
 w_6 =大家 w_7 =觉得 w_8 =他 w_9 =幽默 w_{10} =。

➤ 等价表示方法：one-hot表示法

$|V|$

$\begin{bmatrix} \vdots \end{bmatrix}$



所有的词
按照出现
顺序排序



每个词对
应唯一的
下标

风趣

$\begin{bmatrix} 0 \\ \mathbf{1} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$

幽默

$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{1} \\ 0 \end{bmatrix}$

2. 自然语言处理方法

问题：

风趣 \otimes 幽默

$$\begin{bmatrix} 0 \\ \mathbf{1} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

\times

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \mathbf{1} \\ 0 \end{bmatrix}$$

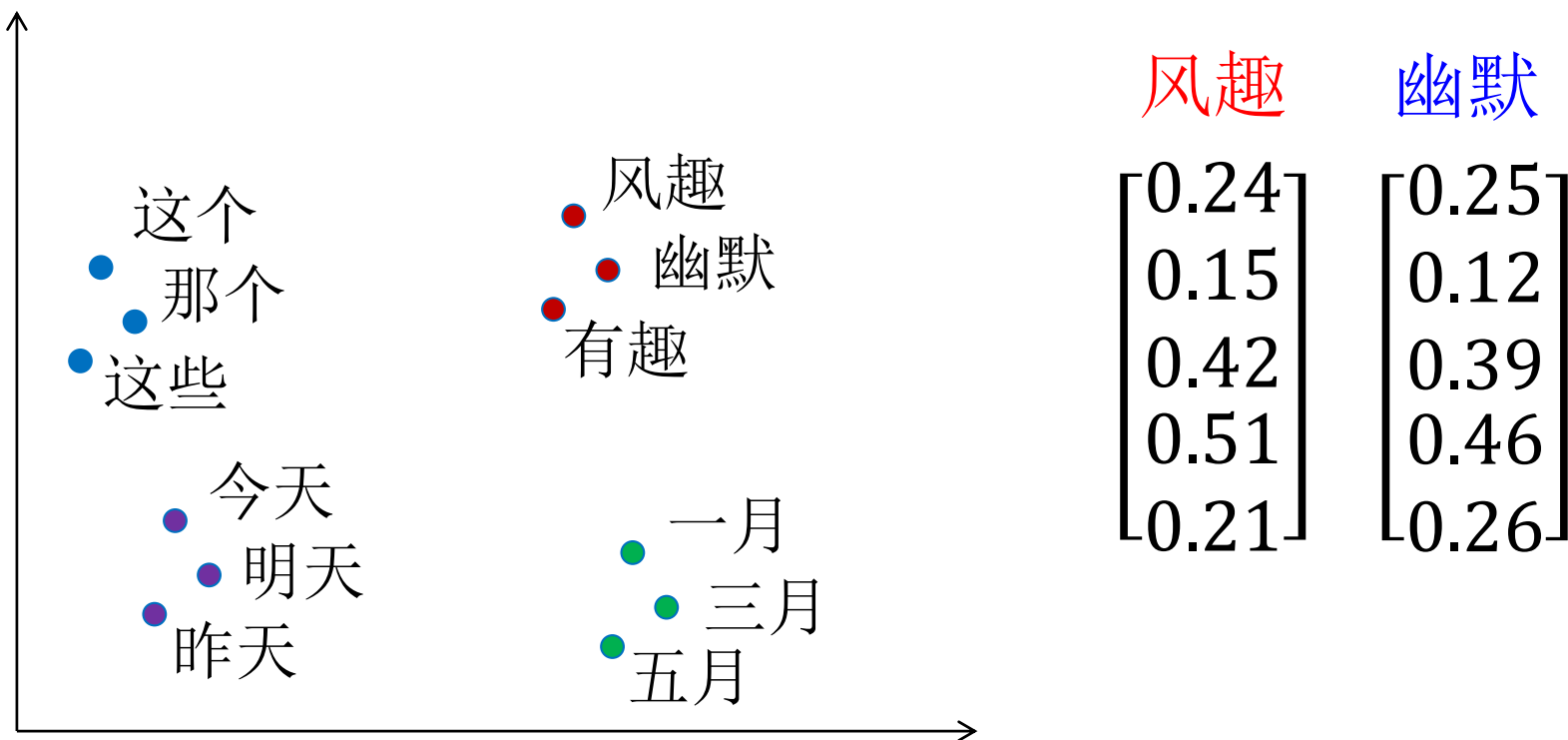
$= 0$



任意两个词之间的相似度都为0！

2. 自然语言处理方法

◆ 分布式表示 — 词向量表示 (word2vec)



低维、稠密的连续实数空间

2. 自然语言处理方法

◆ 基于文本的词向量学习

$$L = \begin{bmatrix} \text{有趣} & \dots & \text{风情} & \text{幽默} \end{bmatrix}_D \quad L \in R^{D \times V}$$

Diagram illustrating the word embedding matrix L . The matrix is shown as a grid of red dots representing word embeddings. The columns are labeled with words: 有趣, ..., 风情, 幽默. The matrix is enclosed in large square brackets with a subscript D . To the right of the matrix, the text $L \in R^{D \times V}$ is displayed.

➤ 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. $V=80000$

2. 自然语言处理方法

很

$p(\text{风趣}|\text{很})$

$p(\text{幽默}|\text{很})$

$$\begin{bmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \end{bmatrix}$$

$$p \left(\begin{bmatrix} 0.25 \\ 0.12 \\ 0.39 \\ 0.46 \\ 0.26 \end{bmatrix} \middle| \begin{bmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \end{bmatrix} \right)$$

$$p \left(\begin{bmatrix} 0.24 \\ 0.15 \\ 0.42 \\ 0.51 \\ 0.21 \end{bmatrix} \middle| \begin{bmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \end{bmatrix} \right)$$

$$f \left(\begin{bmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \\ 0.25 \\ 0.12 \\ 0.39 \\ 0.46 \\ 0.26 \end{bmatrix} \right)$$

很风趣

vs.

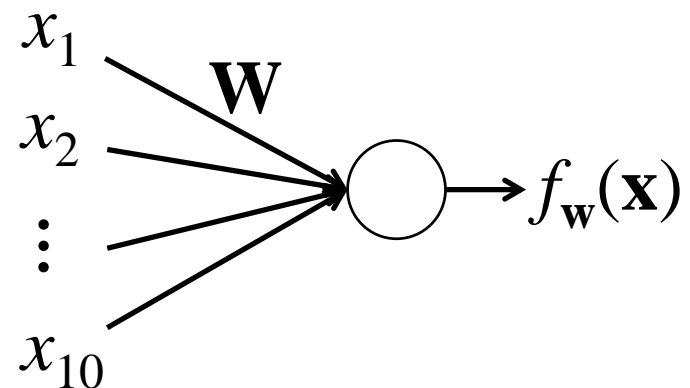
$$f \left(\begin{bmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \\ 0.24 \\ 0.15 \\ 0.42 \\ 0.51 \\ 0.21 \end{bmatrix} \right)$$

很幽默

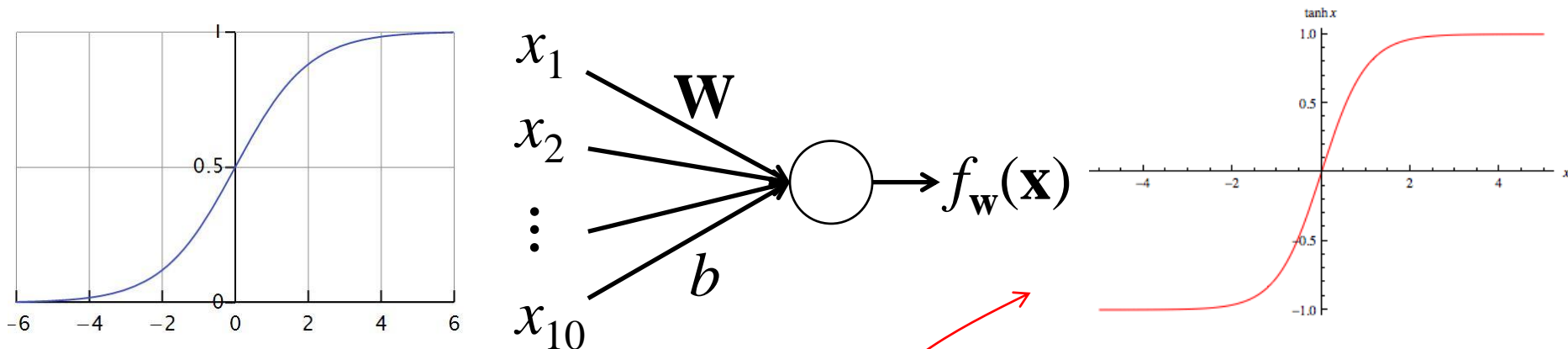
2. 自然语言处理方法

$p(\text{风趣}|\text{很})$

$$f \begin{pmatrix} 0.01 \\ 0.59 \\ 0.18 \\ 0.05 \\ 0.47 \\ 0.25 \\ 0.12 \\ 0.39 \\ 0.46 \\ 0.26 \end{pmatrix} = f \begin{pmatrix} w_1 \times 0.01 \\ w_2 \times 0.59 \\ w_3 \times 0.18 \\ w_4 \times 0.05 \\ w_5 \times 0.47 \\ w_6 \times 0.25 \\ w_7 \times 0.12 \\ w_8 \times 0.39 \\ w_9 \times 0.46 \\ w_{10} \times 0.26 \end{pmatrix} = f(\mathbf{WX})$$



2. 自然语言处理方法



$$h_{W,b}(x) = f(W^T x + b)$$

$$f(z) = \frac{1}{1 + \exp(-z)}$$

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

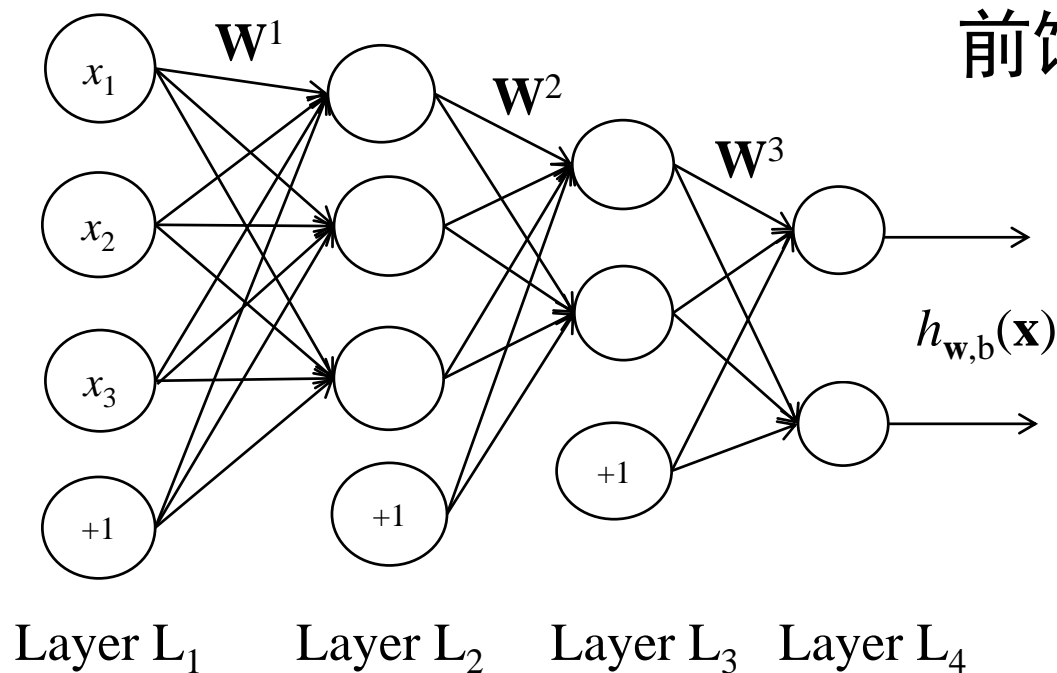
f : 非线性激活函数

$$f'(z) = f(z)(1 - f(z))$$

$$f'(z) = 1 - f^2(z)$$

2. 自然语言处理方法

前馈神经网络



$$f\left(W^3\left(f\left(W^2\left(f\left(W^1x + b^1\right)\right) + b^2\right)\right) + b^3\right)$$

2. 自然语言处理方法

① Input Window

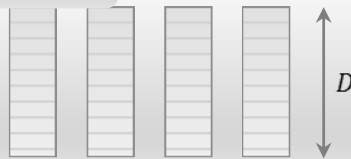
Text 这个人说话很风趣 ...

$p(\text{风趣} \mid \text{这个, 人, 说话, 很})$

② Lookup Table

将每个词通过词向量矩阵L映射为低维实数向量

LT_w



concatenate

这个: (0.2, 0.1); 人: (0.1, 0.3);
说话: (0.4, 0.2); 很: (0.5, 0.4)

③ Linear

$M^1 \times \odot$



H

拼接所有词的向量, 形成一个向量

这个人说话很:
(0.2, 0.1, 0.1, 0.3, 0.4, 0.2, 0.5, 0.4)

$$\begin{pmatrix} 0.1 & 0 & 0.2 & 0.4 & 0.2 & 0.1 & 0 & 0.3 \\ 0.5 & 0.4 & 0.2 & 0 & 0.2 & 0.6 & 0 & 0.2 \end{pmatrix} \times \begin{pmatrix} 0.2 \\ 0.1 \\ 0.1 \\ 0.3 \\ 0.4 \\ 0.2 \\ 0.5 \\ 0.4 \end{pmatrix}$$

$M^1 \times x$



隐藏层:

线性映射+非线性变换

$\begin{pmatrix} 0.38 \\ 0.25 \end{pmatrix}$

2. 自然语言处理方法

① Input Window

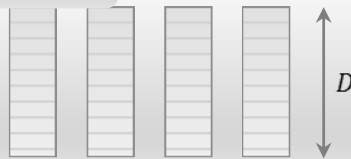
Text 这个人说话很风趣 ...

$p(\text{风趣} \mid \text{这个, 人, 说话, 很})$

② Lookup Table

将每个词通过词向量矩阵 L 映射为低维实数向量

LT_w

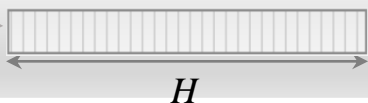


concatenate

这个: (0.2, 0.1); 人: (0.1, 0.3);
说话: (0.4, 0.2); 很: (0.5, 0.4)

③ Linear

$M^1 \times \odot$

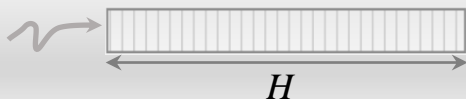


H

拼接所有词的向量，形成一个向量

这个 人 说话 很:
(0.2, 0.1, 0.1, 0.3, 0.4, 0.2, 0.5, 0.4)

④ Tanh



H

隐藏层:

线性映射+非线性变换

$$\mathbf{h}^1 = \tanh \begin{pmatrix} 0.38 \\ 0.25 \end{pmatrix}$$

2. 自然语言处理方法

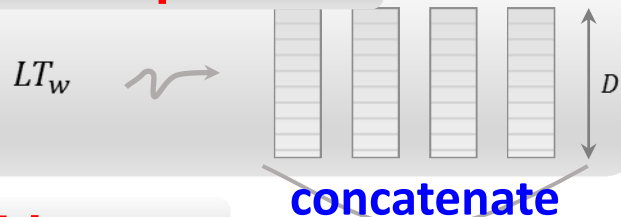
① Input Window

Text 这个人说话很风趣 ...

$p(\text{风趣} \mid \text{这个, 人, 说话, 很})$

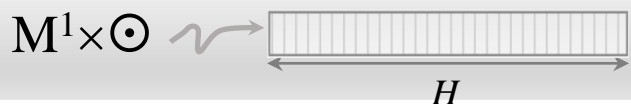
② Lookup Table

将每个词通过词向量矩阵 L 映射为低维实数向量



这个: (0.2, 0.1); 人: (0.1, 0.3);
说话: (0.4, 0.2); 很: (0.5, 0.4)

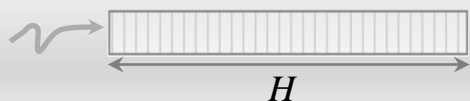
③ Linear



拼接所有词的向量，形成一个向量

这个 人 说话 很:
(0.2, 0.1, 0.1, 0.3, 0.4, 0.2, 0.5, 0.4)

④ Tanh



隐藏层:

线性映射+非线性变换

+ Linear

$M^2 \times \odot$

H

$M^2 \times h$

$\text{softmax}(M^2 \times h)$

2. 自然语言处理方法

◆ 问题

- 仅对小窗口的历史信息进行建模，例如，5-gram 语言模型，仅考虑前面4个词的历史信息

$$p(w_t | w_{t-1} \cdots w_{t-n+1})$$

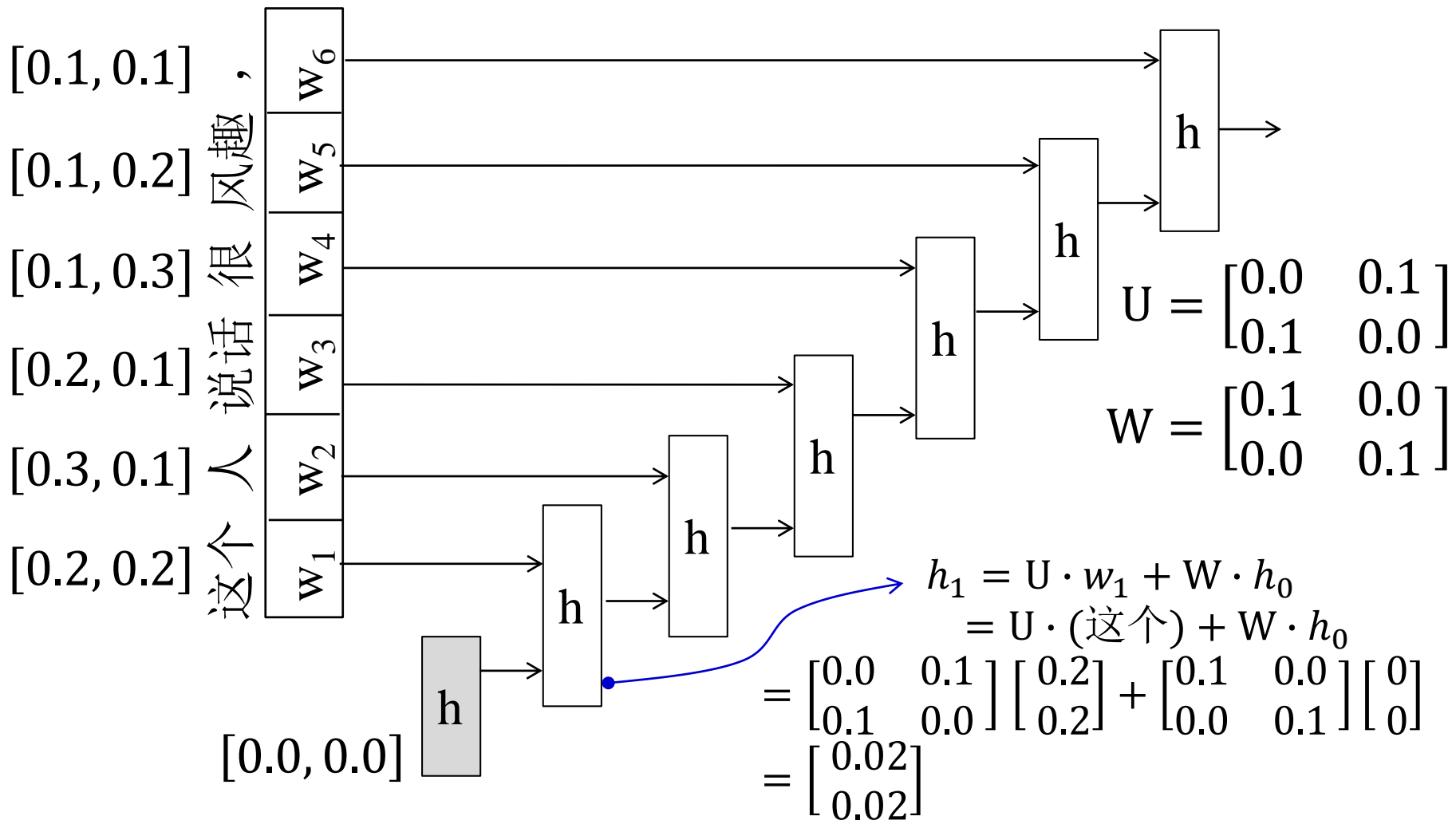


能否对所有的历史信息进行建模，即第 t 个词的语言模型概率依赖于所有前 $t-1$ 个词

$$p(w_t | w_{t-1} \cdots w_2 w_1)$$

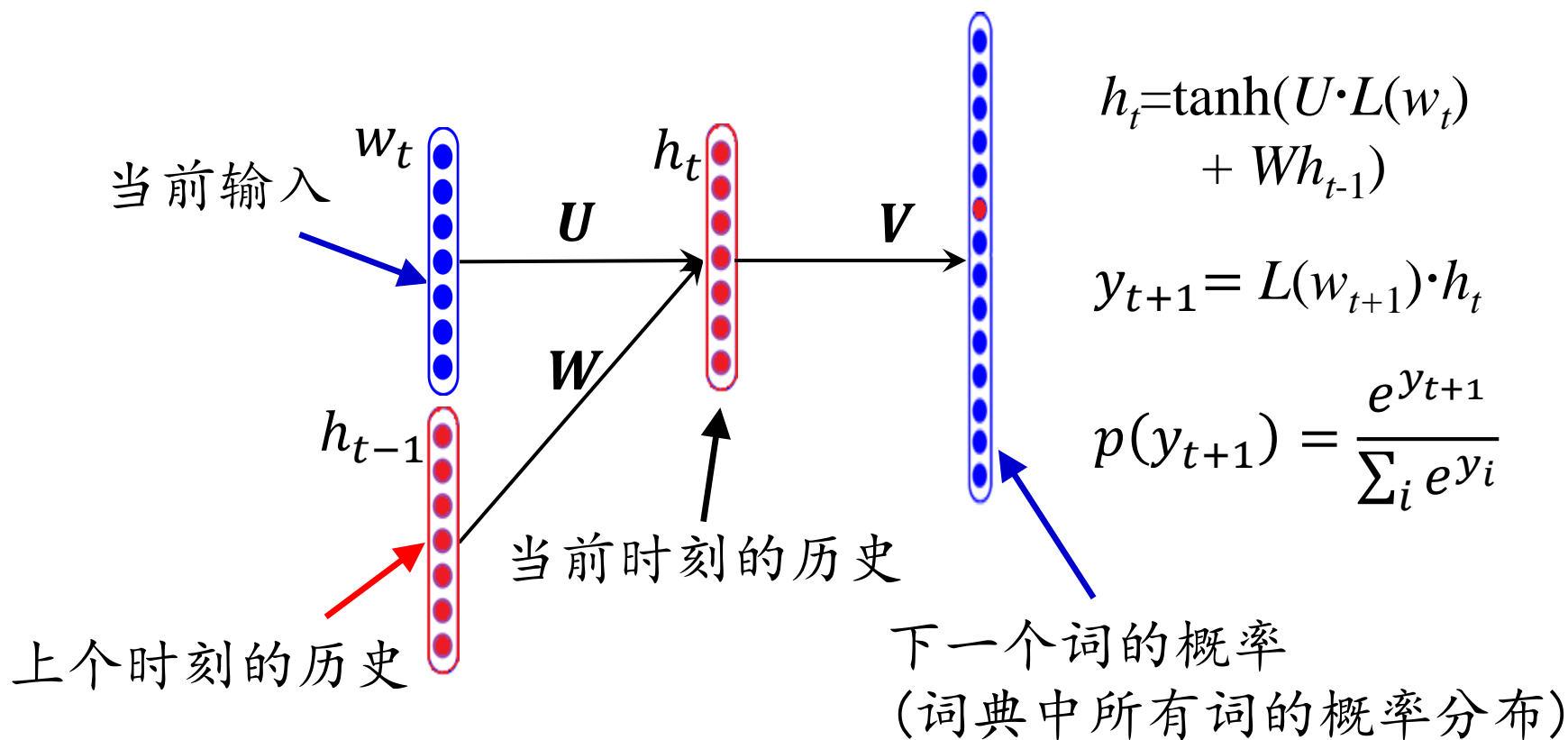
2. 自然语言处理方法

◆ 循环神经网络



2. 自然语言处理方法

- 输入: $t - 1$ 时刻历史 h_{t-1} 与 t 时刻输入 w_t
- 输出: t 时刻历史 h_t 与 下个时刻 $t + 1$ 输入 的概率



2. 自然语言处理方法

问题：

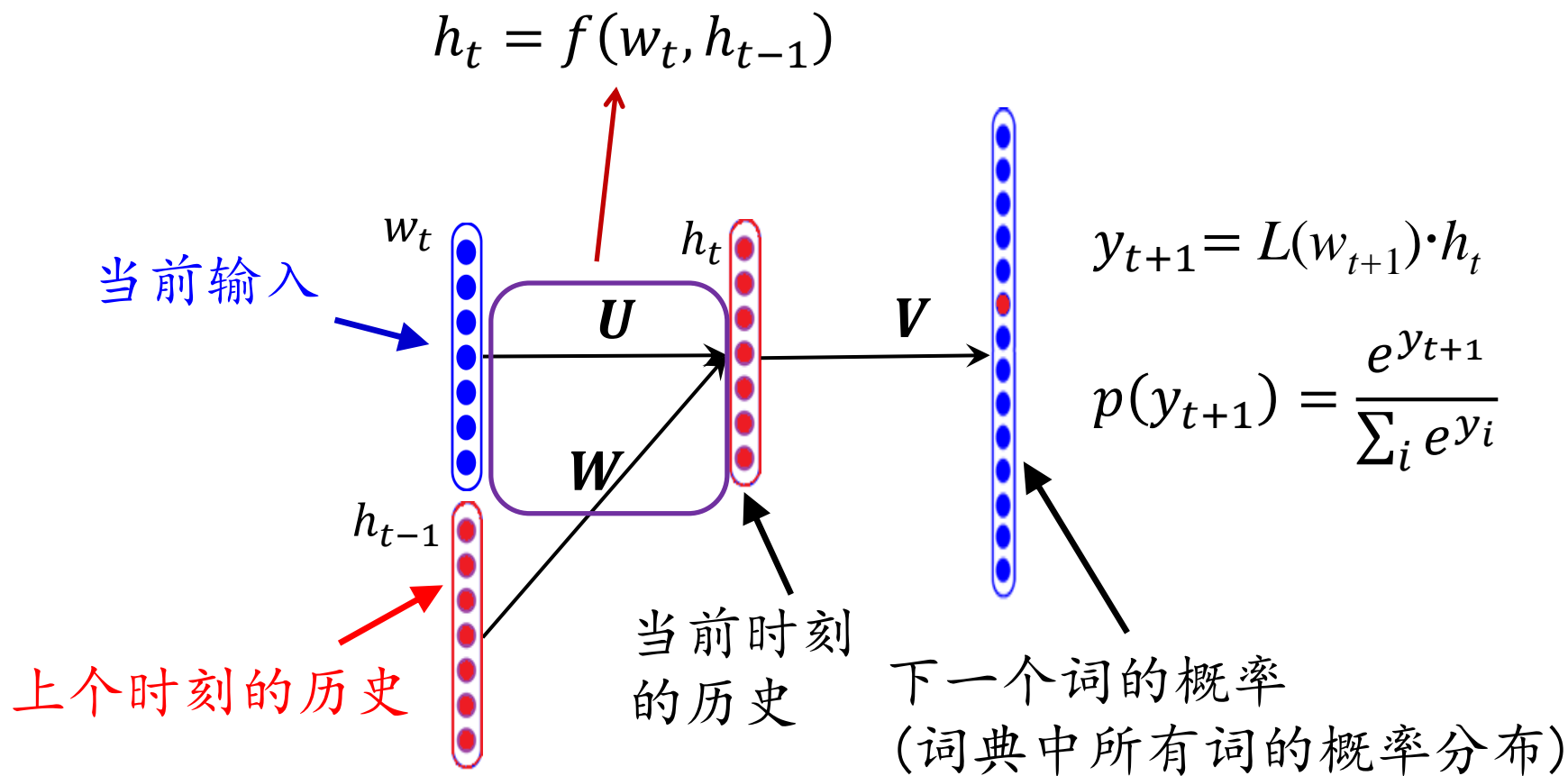
梯度消失和爆炸： 参数 W 经过多次传递后，易发生梯度消失 (<1) 或爆炸 (>1)



有选择地保留和遗忘： 通过某种策略有选择地保留或者遗忘 t 时刻的信息

2. 自然语言处理方法

◆ 长短时记忆网络 (Long/Short Term Memory, LSTM)



2. 自然语言处理方法

LSTM

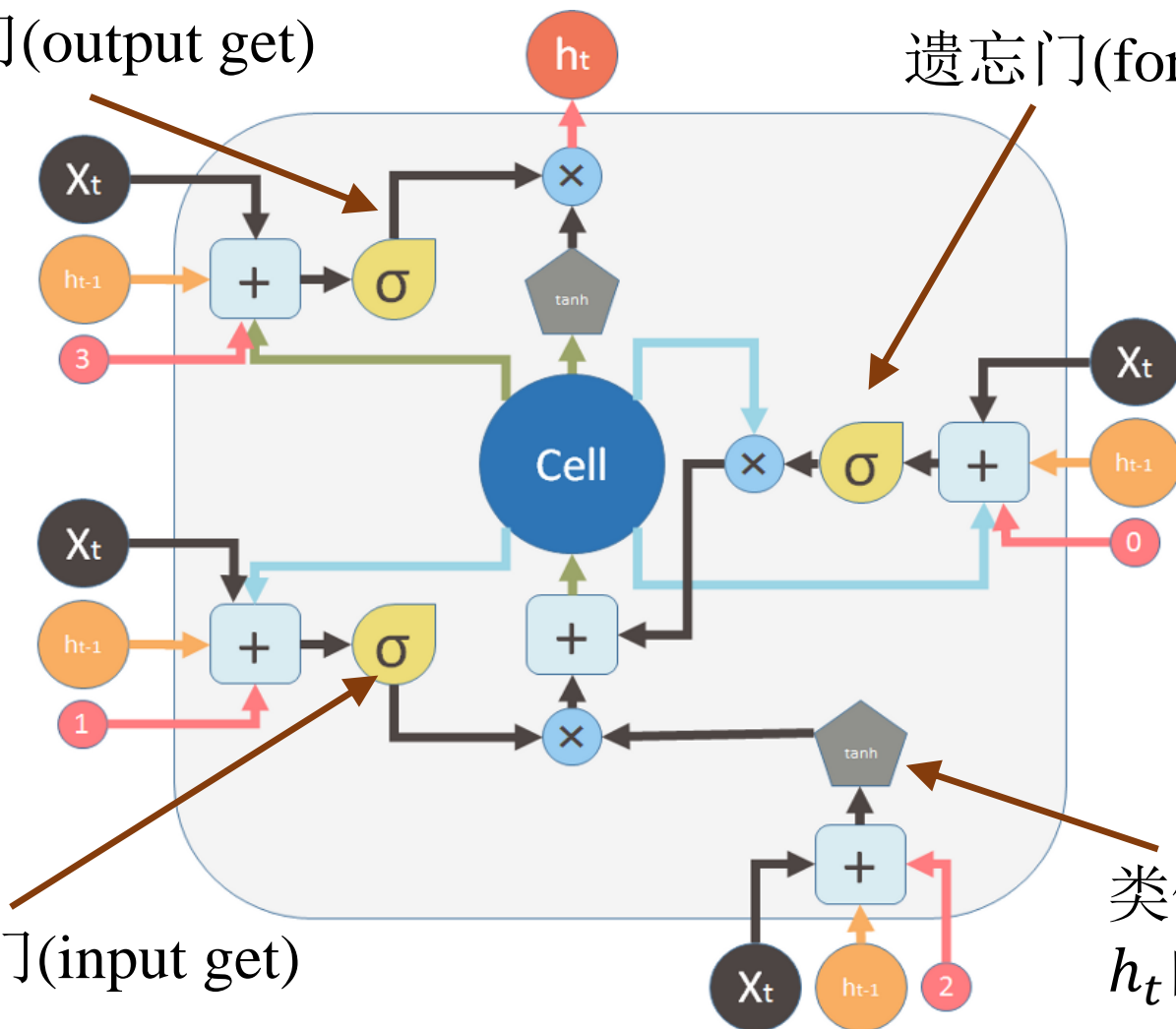
$$h_t = f(w_t, h_{t-1})$$

输出门(output get)

遗忘门(forget get)

输入门(input get)

类似于RNN中
 h_t 的计算



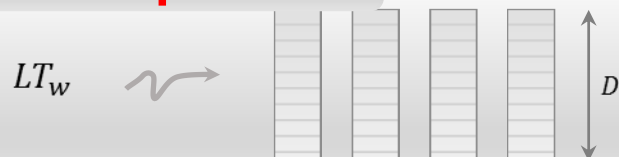
2. 自然语言处理方法

◆ 词向量表示

① Input Window

Text 这个人说话很风趣 ...

② Lookup Table



$p(\text{风趣} \mid \text{这个, 人, 说话, 很})$

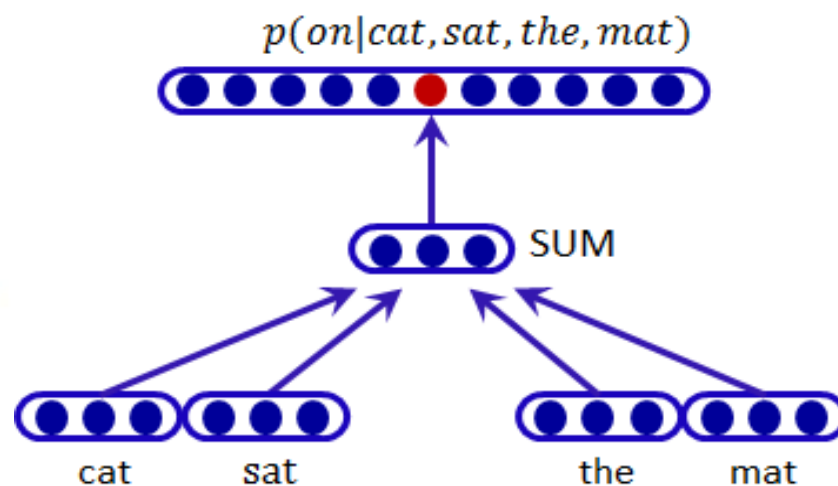
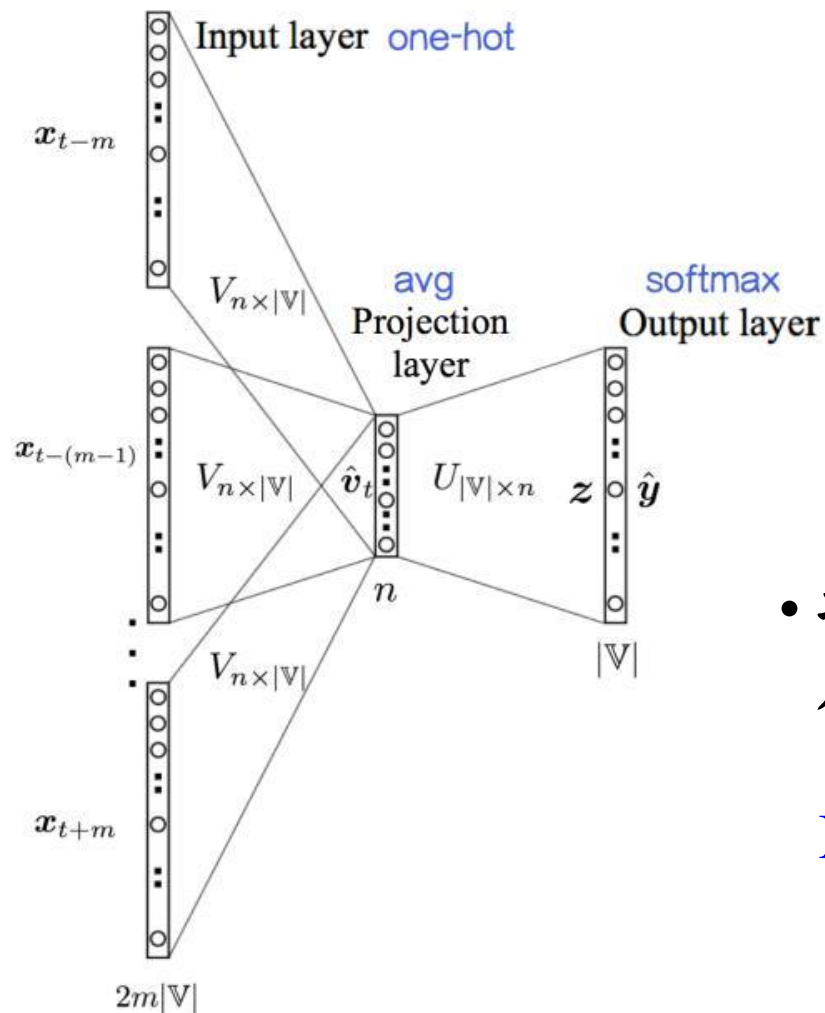


这个: (0.2, 0.1);	人: (0.1, 0.3);
说话: (0.4, 0.2);	很: (0.5, 0.4)

将每个词通过词向量矩阵 L 映射为低维实数向量

2. 自然语言处理方法

①用周围词预测中间词的方法 — 连续词包模型(CBOW)

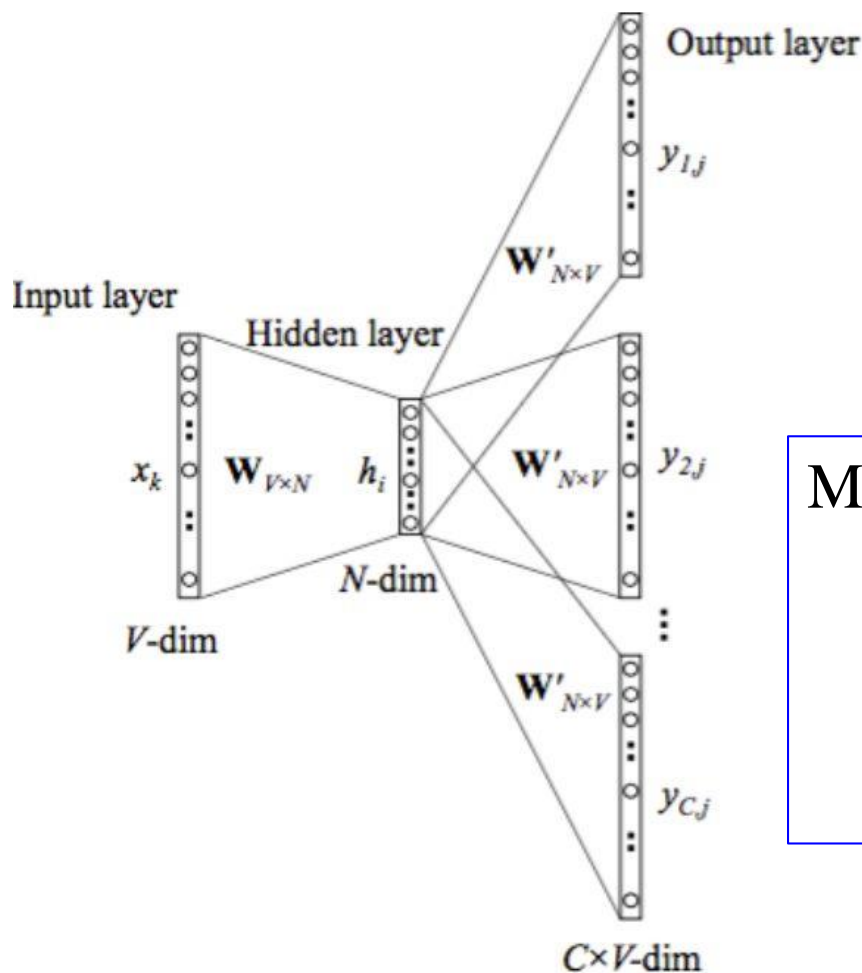


- 将相邻的词向量直接相加得到隐层，用隐层预测中间词的概率

Mikolov et al. (2013)

2. 自然语言处理方法

② 连续 skip-gram 模型：通过中间词预测周围词的概率




Mikolov, Tomas, K. Chen et al .
2013. Efficient estimation of
word representation in vector
space. arXiv preprint arXiv:
1301.3781, 2013

2. 自然语言处理方法

➤ 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/>
-

内容提要

1. 问题的提出
2. 自然语言处理方法
-  3. 应用举例
4. 技术现状
5. 结束语

3. 应用举例

- ◆ 汉语自动分词
- ◆ 机器翻译
- ◆ 问答/对话系统
- ◆ CASIA 相关工作

3. 应用举例

◆ 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



3. 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！

3. 应用举例

➤ 最大匹配法 (基于规则或模板的方法)

他是研究生。

↑ — 6 — → |

↑ — 5 — → |

↑ — 4 — → |

...

他/ 是研究生。

↑ — — — → |

...

他/ 是/ 研究生。

...

他/ 是/ 研究生/ 。



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)

3. 应用举例

➤ 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 $S = z_1 z_2 \dots z_m$, 假设 $W = w_1 w_2 \dots w_k$ ($1 \leq k \leq n$) 是一种可能的切分结果。那么,

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

最基本的做法是以词为独立的统计基元, 但效果不佳。

3. 应用举例

➤ 基于隐马尔可夫模型(HMM)的分词方法 (传统的统计方法, 生成式)

HMM: $\mu = (A, B, \pi)$

- (1)初始状态; (2)状态转移概率;
- (3)观察概率; (4)状态个数; (5)输出数目

思路:

如果把汉语自动分词结果作为观察序列 $O=O_1O_2...O_T$, 那么, 我们要求解的是: $\hat{O} = \arg \max_O p(O|\mu)$ 。对于词性标注而言, 则需求解: $\hat{Q} = \arg \max_Q p(Q|O, \mu)$ 。

3. 应用举例

进一步解释:

- (1) 估计HMM模型 $\mu=(A, B, \pi)$ 的参数;
- (2) 对于任意给定的一个输入句子及其可能的输出序列 O , 求找所有可能的 O 中使概率 $p(O|\mu)$ 最大的解;
- (3) 快速地选择“最优”的状态序列(词性序列), 使其最好地解释观察序列。

3. 应用举例

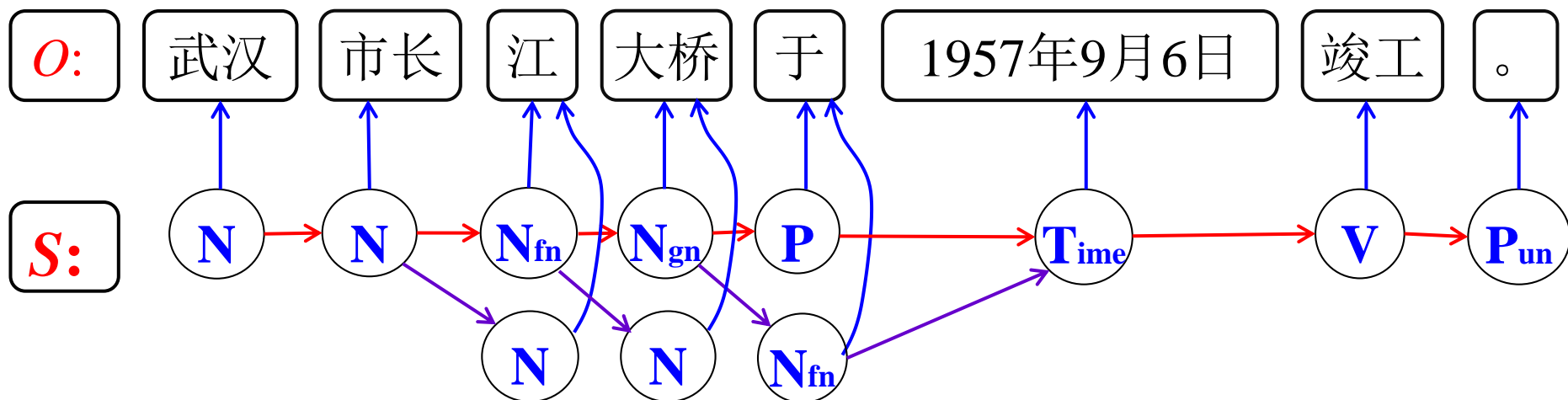
例如：

武汉市长江大桥于1957年9月6日竣工。

列出所有可能的切分：

- ① 武汉市/N 长江/N 大桥/N 于/P 1957年9月6日
/Time 竣工/V。 /Pun
- ② 武汉/N 市长/N 江大桥/N 于/P 1957年9月6日
/Time 竣工/V。 /Pun

3. 应用举例



- 武汉/N 市长/N 江/N_{fn} 大桥/N_{gn} 于/P 1957年9月6日/Time 竣工/V 。/Pun
- 武汉/N 市长/N 江/N 大桥/N_{gn} 于/P 1957年9月6日/Time 竣工/V 。/Pun
- 武汉/N 市长/N 江/N_{fn} 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun
- 武汉/N 市长/N 江/N 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun
- 武汉/N 市长/N 江/N_{fn} 大桥/N 于/N_{fn} 1957年9月6日/Time 竣工/V 。/Pun
-

3. 应用举例

➤ 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/ 。/

3. 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

.....

3. 应用举例

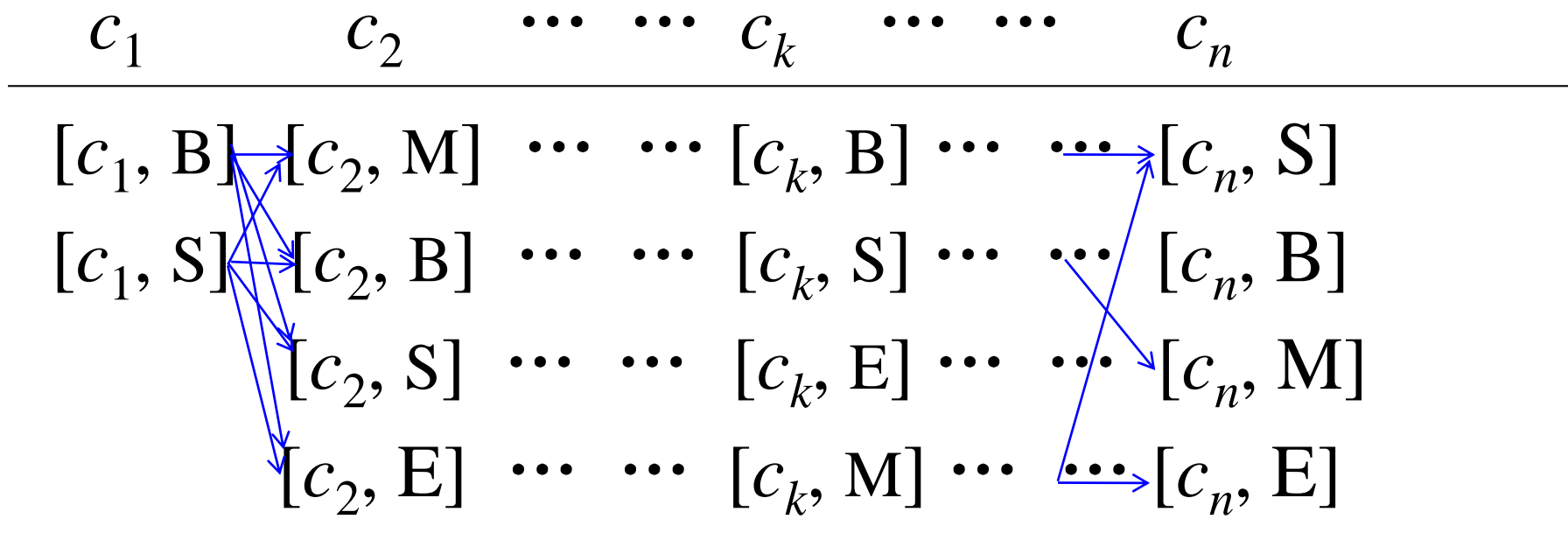
● 问题分析：

基于字的区分模型有利于处理集外词，而基于词的生成模型更多地考虑了词汇之间以及词汇内部字与字之间的依存关系。因此，可以将两者的优势结合起来。

➤ 生成式与区分式组合的方法 (传统的统计方法)

✧ 结合方法1：将待切分字符串的每个汉字用 $[c, t]_i$ 替代，以 $[c, t]_I$ 作为基元，利用语言模型选取全局最优(生成式模型)。

3. 应用举例



[上, B] [海, E] [计, B] [划, E] [到, S] [本, S] [世, B] [纪, E] \dots

$$P([c, t]_1^n) \approx \prod_{i=1}^n P([c, t]_i | [c, t]_{i-k}^{i-1})$$

3. 应用举例

实验结果:

- ✧ 利用第二届 SIGHAN Bakeoff 评测语料(2005)
- ✧ 4种语料: 北大、台湾中研院、香港城大、微软
- ✧ 分词正确率(P):
 - (1) 基于词的 3-gram: $P=89.8\%$
 - (2) 基于字的 CRF: $P=94.3\%$
 - (3) 融合方法 3-gram: $P=95.0\%$

K. Wang, C. Zong, and K. Su. Which is More Suitable for Chinese Word Segmentation, the Generative Model or the Discriminative One? In *Proc. of PACLIC-23*. 3-5 Dec. 3-5, 2009, HK. pp. 827-834

3. 应用举例

➤ 分析:

- 该方法的优点:

- (1) 充分考虑了相邻字之间的依存关系进行建模;
- (2) 相对于区分模型, 对集内词(IV)有较好的鲁棒性。

- 弱点:

难以利用后续的上下文信息。

- 回顾—基于字的区分式模型的优点:

- (1) 与基于词的方法相比, 对集外词(OOV)具有更好的鲁棒性;
- (2) 相对于生成模型, 容易处理更多的特征。

3. 应用举例

✧ 结合方法2：插值法把两种方法结合起来

$$Score(t_k) = \alpha \times \log(P([c, t]_k \mid [c, t]_{k-2}^{k-1})) + (1 - \alpha) \times \log(P(t_k \mid c_{k-2}^{k+2}))$$

(0.0 ≤ α ≤ 1.0)

Generative score

Discriminative score

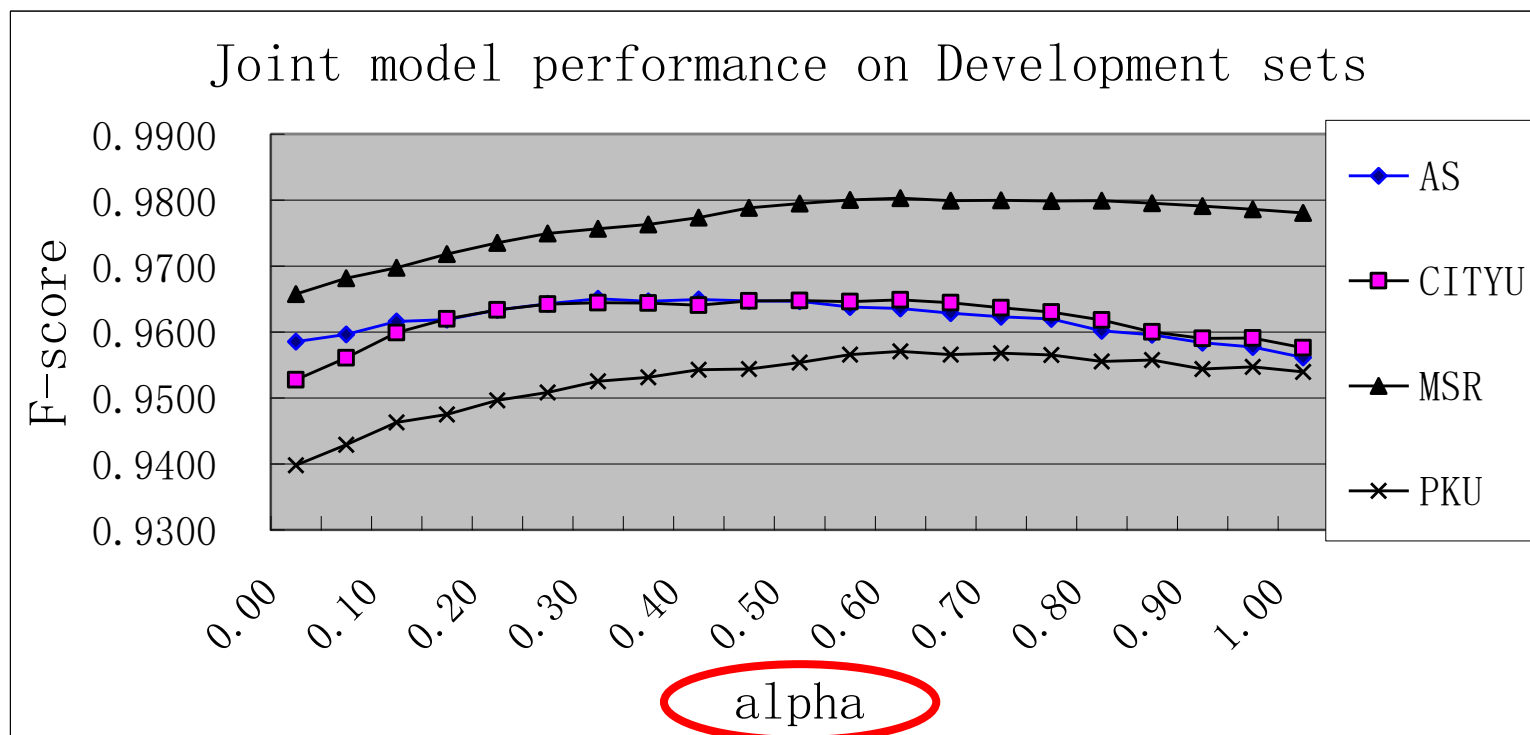
● 这样做的优点：

充分结合了基于字的生成模型和基于字的区分式模型的优点。

3. 应用举例

性能测试

语料：2005 年 SIGHAN Bakeoff 语料，取少量做开发集。



3. 应用举例

Corpus	Model	R	P	F	R_{Oov}	R_{IV}
AS	Generative	0.958	0.938	0.948	0.518	0.978
	Discriminative	0.955	0.946	0.951	0.707	0.967
	Joint	0.962	0.950	0.956	0.679	0.975
CITYU	Generative	0.951	0.937	0.944	0.609	0.978
	Discriminative	0.941	0.944	0.942	0.708	0.959
	Joint	0.957	0.951	0.954	0.691	0.979
MSR	Generative	0.974	0.967	0.970	0.561	0.985
	Discriminative	0.957	0.962	0.960	0.719	0.964
	Joint	0.974	0.971	0.972	0.659	0.983
PKU unconverted (ucvt.) case	Generative	0.929	0.933	0.931	0.435	0.959
	Discriminative	0.922	0.941	0.932	0.620	0.941
	Joint	0.935	0.946	0.941	0.561	0.958

3. 应用举例

Corpus	Model	R	P	F	R_{OOV}	R_{IV}
PKU converted (cvt.) case	Generative	0.952	0.951	0.952	0.503	0.968
	Discriminative	0.940	0.951	0.946	0.685	0.949
	Joint	0.954	0.958	0.956	0.616	0.966
Overall	Generative	0.953	0.946	0.950	0.511	0.973
	Discriminative	0.944	0.950	0.947	0.680	0.956
	Joint	0.957	0.955	0.956	0.633	0.971

总体性能：相对错误率比区分式模型减少 21%，比生成式模型减少14%。

注：‘(cvt.) case’指已将测试集中的数字、西文字母等编码转换，使其与训练集中的编码一致，‘(ucvt.) case’指未做转换。

3. 应用举例

2010 CIPS-SIGHAN 评测结果：

Domains	Mark	OOV Rate	R	P	$F1$	R_{OOV}	R_{IV}
Literature	A	0.069	0.937	0.937	0.937	0.652	0.958
Computer	B	0.152	0.941	0.940	0.940	0.757	0.974
Medicine	C	0.110	0.930	0.917	0.923	0.674	0.961
Finance	D	0.087	0.957	0.956	0.957	0.813	0.971

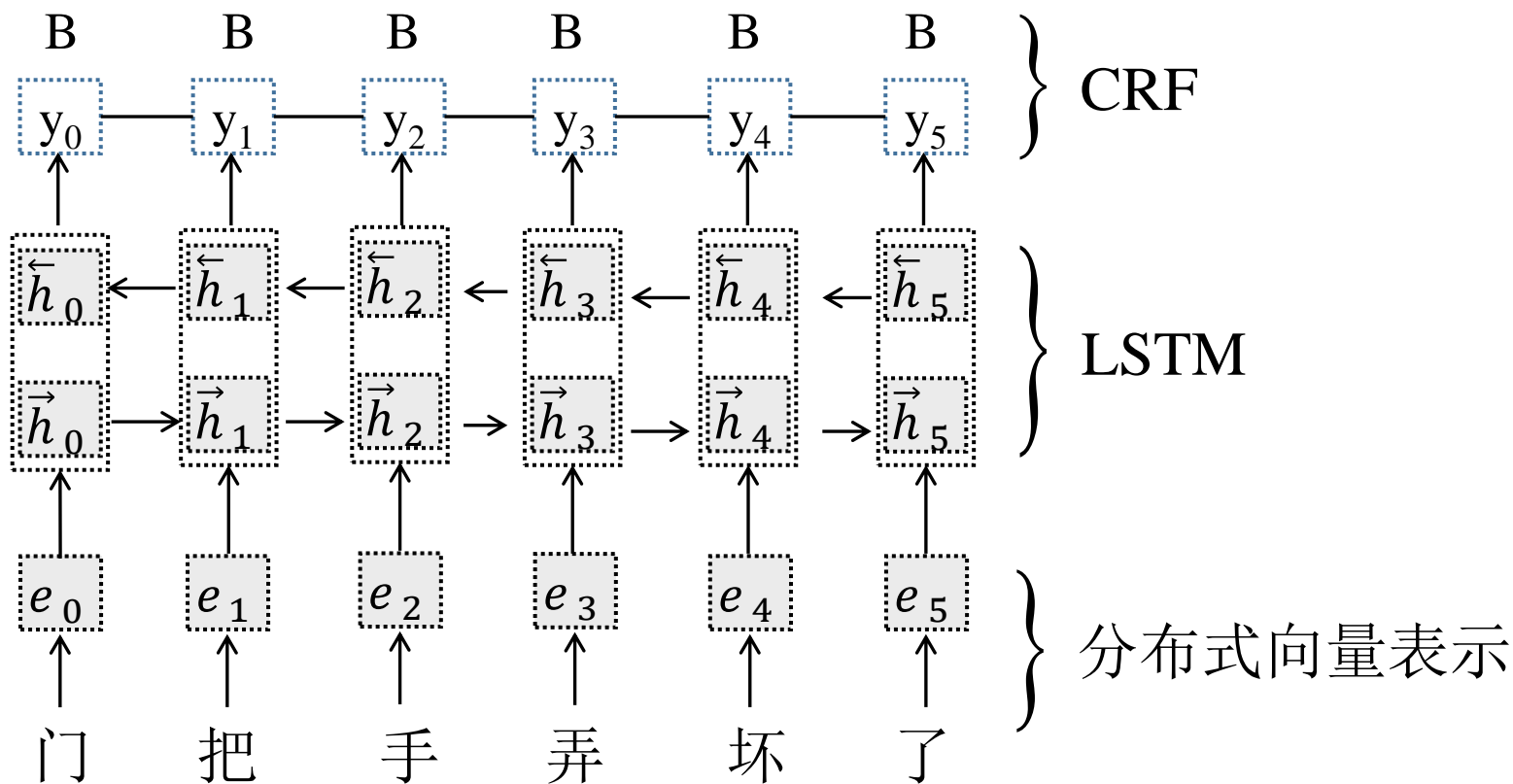
欢迎参阅：

- [1] K. Wang et al. A Character-Based Joint Model for Chinese Word Segmentation. *Proc. COLING 2010*, Aug. 23-27, 2010, pp. 1173-1181
- [2] K. Wang et al. A Character-Based Joint Model for CIPS-SIGHAN Word Segmentation Bakeoff 2010. *Proc. CLP2010*, 2010, pp. 245-248
- [3] K. Wang et al. Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. *ACM TALLIP*, Vol. 11, No.2, 2012

Urheen 分词系统： <http://www.nlpr.ia.ac.cn/cip/software.htm>

3. 应用举例

► 基于神经网络的分词方法



3. 应用举例

➤ 技术现状分析

● 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）。

3. 应用举例

● 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。

3. 应用举例

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）	

3. 应用举例

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

3. 应用举例

- 以古文文本分词为例

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%

3. 应用举例

➤ 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

3. 应用举例

➤ 采用类似的方法可以完成一系列相关任务

● 基本名词短语(base NP)、基本动词短语(base VP)的识别

	COL: 0	COL: 1	TAG	
POS:-4	He	PRP	B-NP	
POS:-3	reckons	VBZ	B-VP	
POS:-2	the	DT	B-NP	Feature Sets
POS:-1	current	JJ	I-NP	
POS: 0	deficit	NN	I-NP	Eestimated TAG
POS:+1	will	MD	B-VP	
POS:+2	narrow	VB	I-NP	
POS:+3	to	TO	B-PP	

the current deficit will...

↑ ↑ ↑ ↑

B-NP I-NP I-NP B-VP

3. 应用举例

● 命名实体(named entity)识别

坐落于江苏省南京市玄武湖公园内的夏璞墩是晋代著名的文学家、科学家夏璞的衣冠冢。



坐落/ 于/ 江苏省/ 南京市/ 玄武湖/ 公园/ 内/ 的/ 夏璞/ 墩/ 是/ 晋代/ 著名/ 的/ 文学家/、/ 科学家/ 夏璞/ 的/ 衣冠冢/ 。

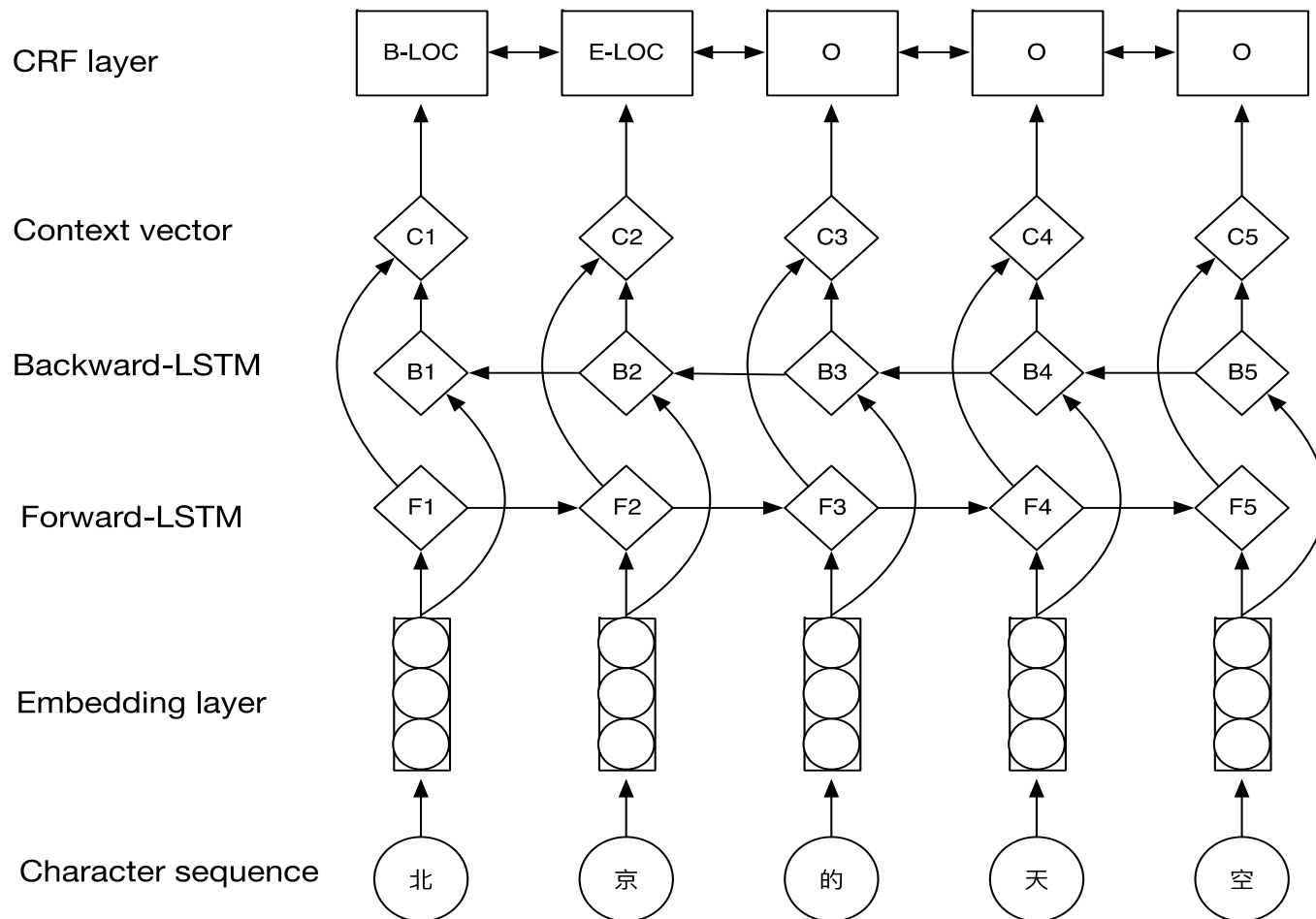


坐落/ 于/ 江苏省/LOC-B 南京市/LOC-I 玄武湖/ORG-B 公园/ ORG-I 内/ 的/ 夏璞/NAM-B 墩/NAM-O 是/ 晋代/ 著名/ 的/ 文学家/、/ 科学家/ 夏璞/NAM-B 的/NAM-O 衣冠冢/ 。

也可以以字为单位

3. 应用举例

● 命名实体(named entity)识别



3. 应用举例

● 语义角色标注(semantic role labelling, SRL)

[他们]_{Agent} [昨天]_{Time} [在北京]_{Location} [讨论]_{Pred} 了 [方案]_{Patient}。

ARG0 ARGM-TMP ARGM-LOC PRED ARG1

序列标注方法：

句子	警察	已	到现场	，	正在	详细	调查	事故	原因
语块	[NP]	[ADVP]	[VP]		[ADVP]	[ADVP]	[VP]	[NP]	[NP]
序列	B-A0	O	O		B-AM-TMP	B-AM-MNR	B-V	B-A1	I-A1
角色	[A0]				[AM-TMP]	[AM-MNR]	[V]	[A1]	

3. 应用举例

● 词义消歧(word sense disambiguation, WSD)

- | | |
|----------------|---------------|
| (1)他打鼓很在行。 | (9) 她会用毛线打毛衣。 |
| (2) 他会打家具。 | (10) 他用尺子打个格。 |
| (3) 他把碗打碎了。 | (11) 他打开了箱子盖。 |
| (4) 他在学校打架了。 | (12) 她打着伞走了。 |
| (5) 他很会与人打交道。 | (13) 他打来了电话。 |
| (6) 他用土打了一堵墙。 | (14) 他打了两瓶水。 |
| (7) 用面打浆糊贴对联。 | (15) 他想打车票回家。 |
| (8) 他打铺盖卷儿走人了。 | (16) 他以打鱼为生。 |
| | |

25+2个不同的含义！

3. 应用举例

基本思路：

每个词表达不同的含意时其上下文（语境）往往不同，也就是说，不同的词义对应不同的上下文，因此，如果能够将多义词的上下文区别开，其词义自然就明确了。

他/P 很/D 会/V 与/C 人/N 打/V 交道/N 。/PU

... -2 -1 ↑ +1 +2 ...

0

基本的上下文信息：词、词性、位置、……

- 特征表示与特征选择
- 分类器： Bayes, ME, SVM, CRFs, ……

3. 应用举例

◆ 汉语自动分词

◆ 机器翻译

◆ 问答/对话系统

◆ CASIA 相关工作

3. 应用举例

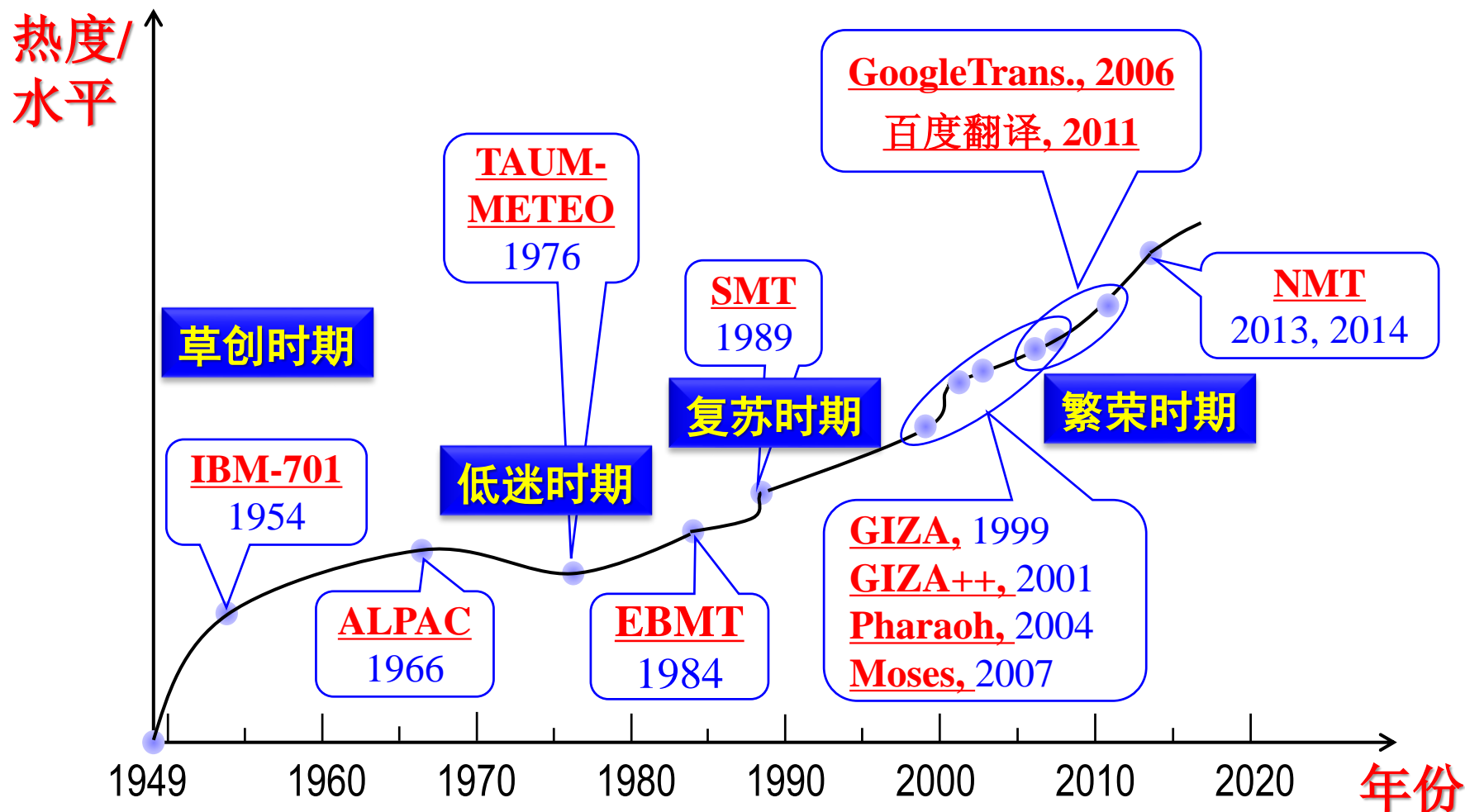
◆ 机器翻译概念

机器翻译 (machine translation, MT) 是用计算机把一种语言(源语言, source language) 翻译成另外一种语言(目标语言, target language) 的学科和技术。



3. 应用举例

➤ 机器翻译技术的发展



3. 应用举例

➤ 机器翻译方法

- 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

3. 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？

3. 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

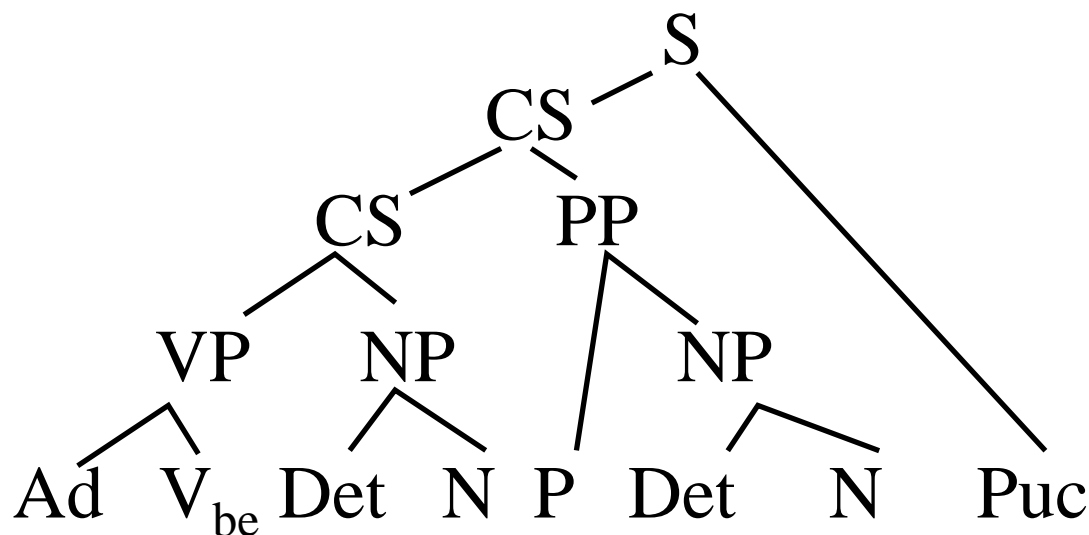
3. 应用举例

给定英语句子：There is a book on the desk.
将其翻译成汉语。

(a)对英语句子进行词法分析

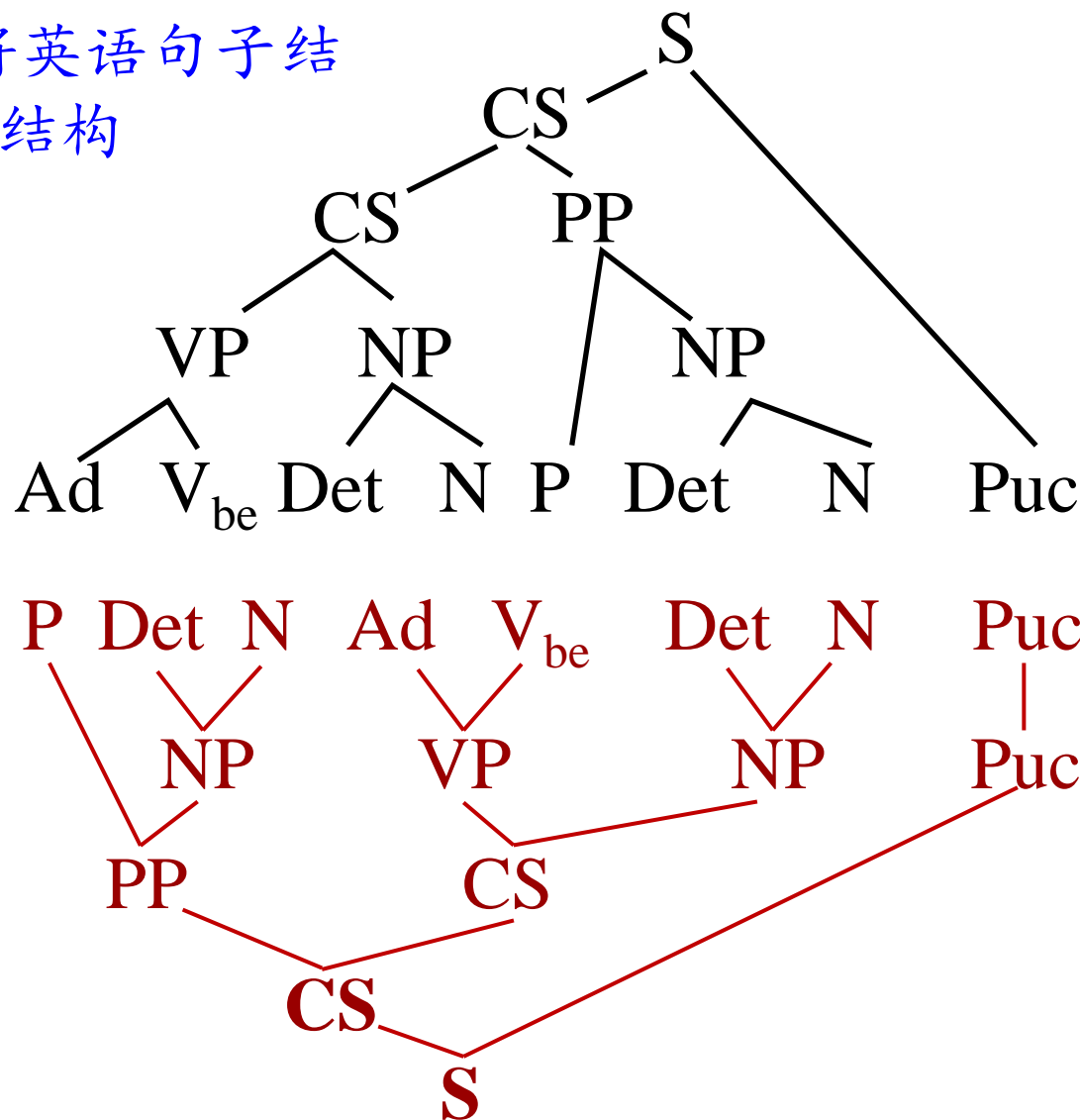
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

(b)对英语句子进行句法结构分析



3. 应用举例

(3) 利用转换规则将英语句子结构转换成汉语句子结构



3. 应用举例

(4)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

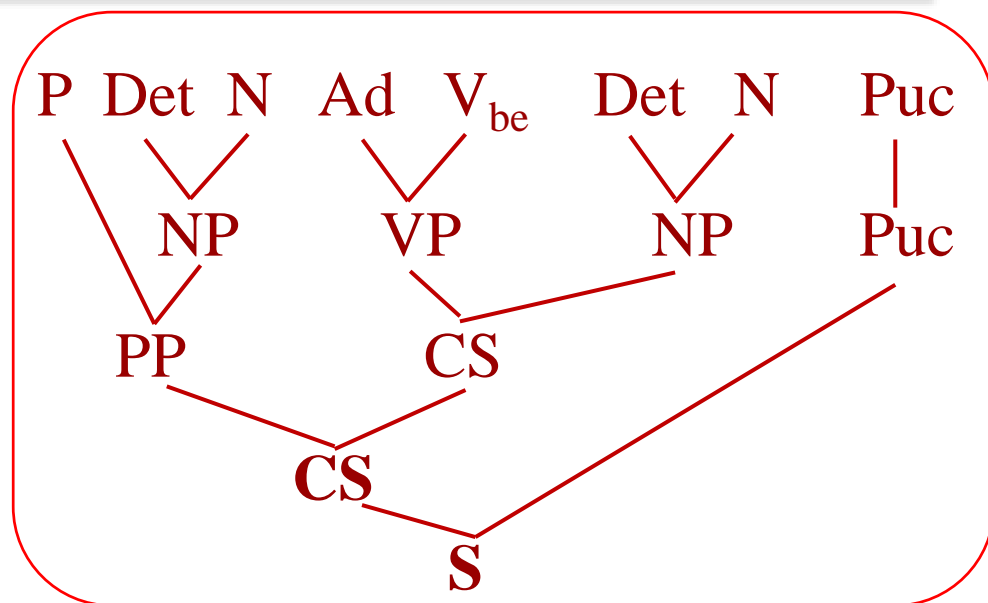
#a, Det, 一

#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

3. 应用举例

方法评价：

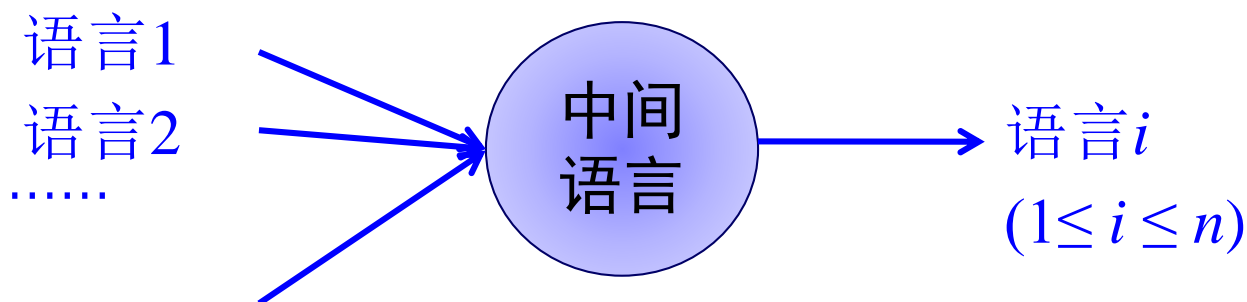
优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

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。

3. 应用举例

- 基于中间语言的翻译方法 (interlingua-based MT)

- 方法：输入语句→中间语言→ 翻译结果
- 代表系统：JANUS (CMU) 早期版本
 - ★ 源语言解析器
 - ★ 比较准确的中间语言(Interlingua)
 - ★ 目标语言生成器(Target Language Generator)

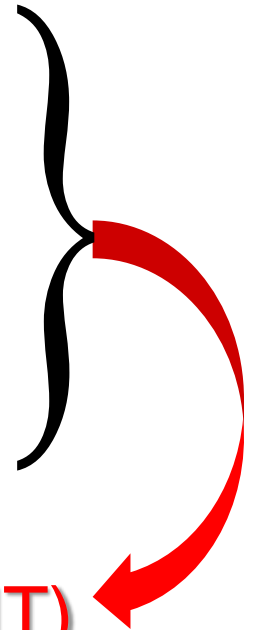


3. 应用举例

- 基于语料库的翻译方法 (corpus-based MT)

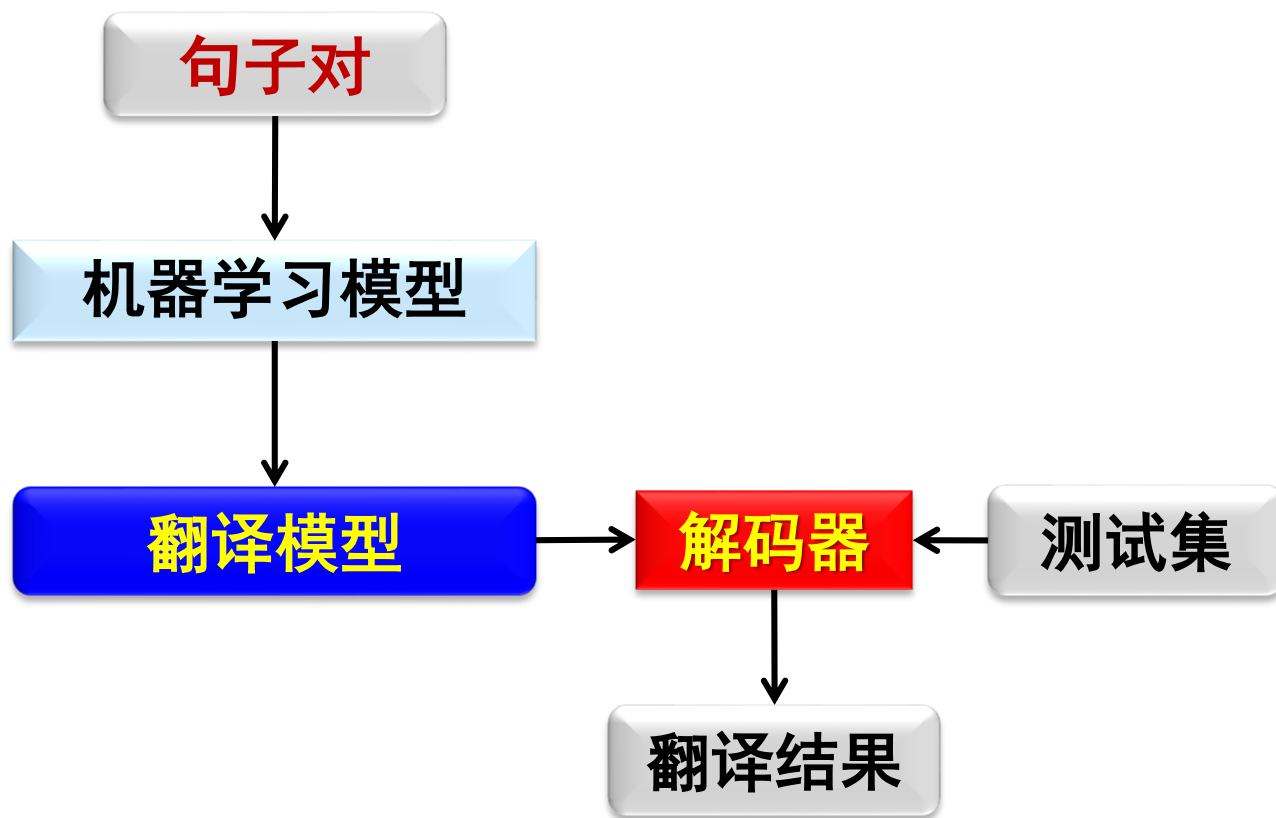
- 基于事例的机器翻译
(example-based machine translation, EBMT, 1984)
- 统计机器翻译
(statistical MT, SMT, 1989)
- 神经网络机器翻译
(neural network MT, NNMT, 2013, 2014)

数据驱动的机器翻译(data-driven MT)



3. 应用举例

数据驱动方法的基本架构：



3. 应用举例

平行句对样本：

merkezdiki dölet apparatliri bilen jaylardiki dölet apparatlrining xizmet hoquqi merkeznning 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 üzishni men'i qilidu, milletler ittipaqliqini buzidighan we milliy bölgünychilik 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。

3. 应用举例

- 统计机器翻译 (statistical machine translation, SMT)

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: $P(C | E) = \frac{P(C) \times P(E | C)}{P(E)}$

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

语言模型

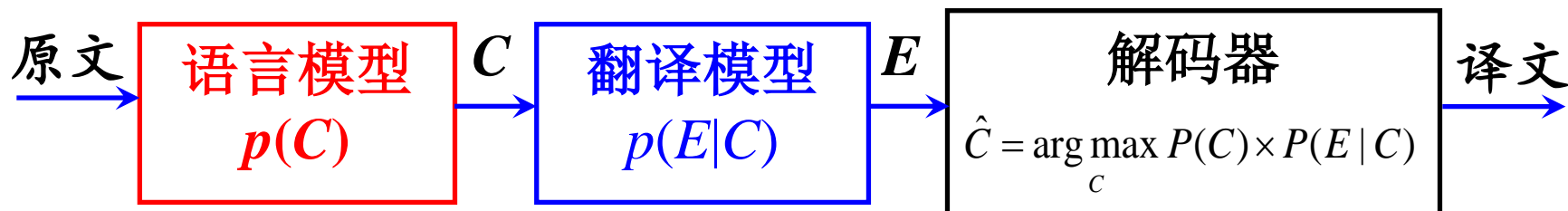
(Language model, LM)

翻译模型

(Translation model, TM)

3. 应用举例

构建解码器(decoder), 快速搜索最优翻译候选:



■ 三个关键问题:

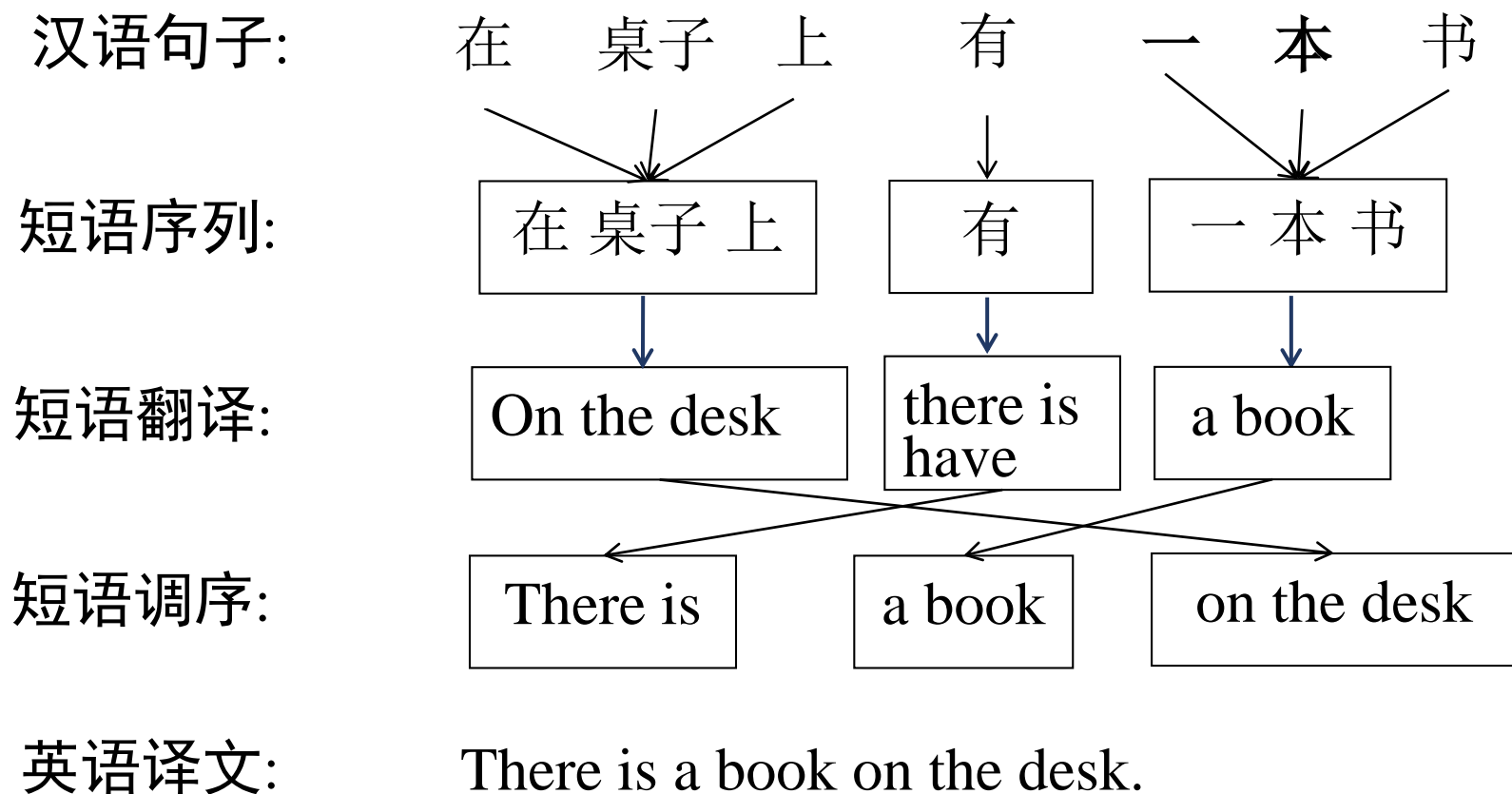
- 估计语言模型概率 $p(C)$;
- 估计翻译模型概率 $p(E|C)$;
- 快速有效地搜索候选译文 C , 使 $p(C) \times p(E|C)$ 最大。

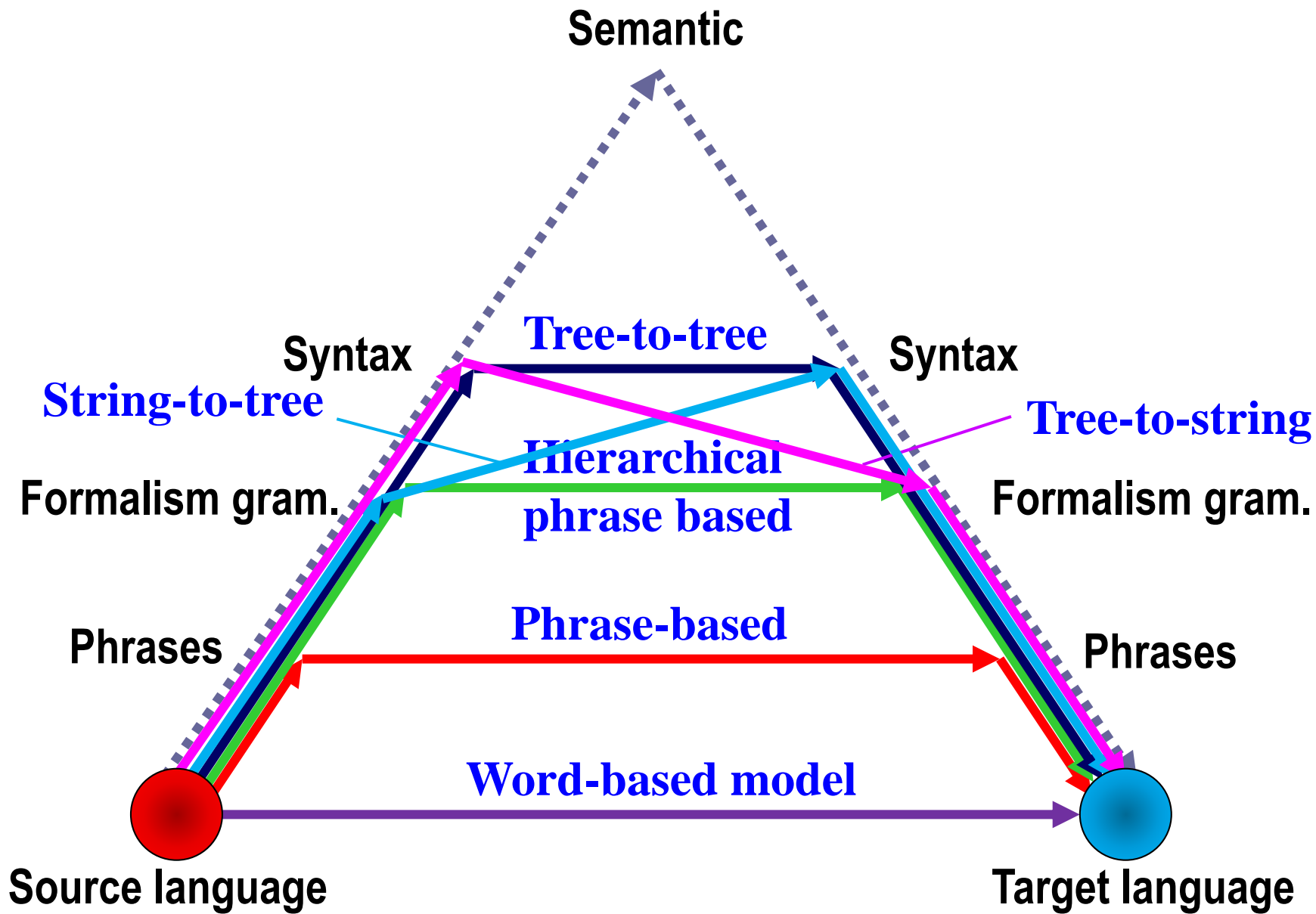
■ 主要任务:

- 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

3. 应用举例

基于短语的翻译模型 (phrase-based translation model):





3. 应用举例

- 神经机器翻译方法

- 基于神经网络的机器翻译

Machine Translation Based on Neural Network

- 神经机器翻译

Neural Machine Translation, NMT



Yoshua Bengio

Montreal Institute for Learning
Algorithms (MILA)



Kyunghyun Cho

New York University (NYU)

3. 应用举例

给定一个源语言句子: $C = c_1^l \equiv c_1 c_2 \cdots c_l$

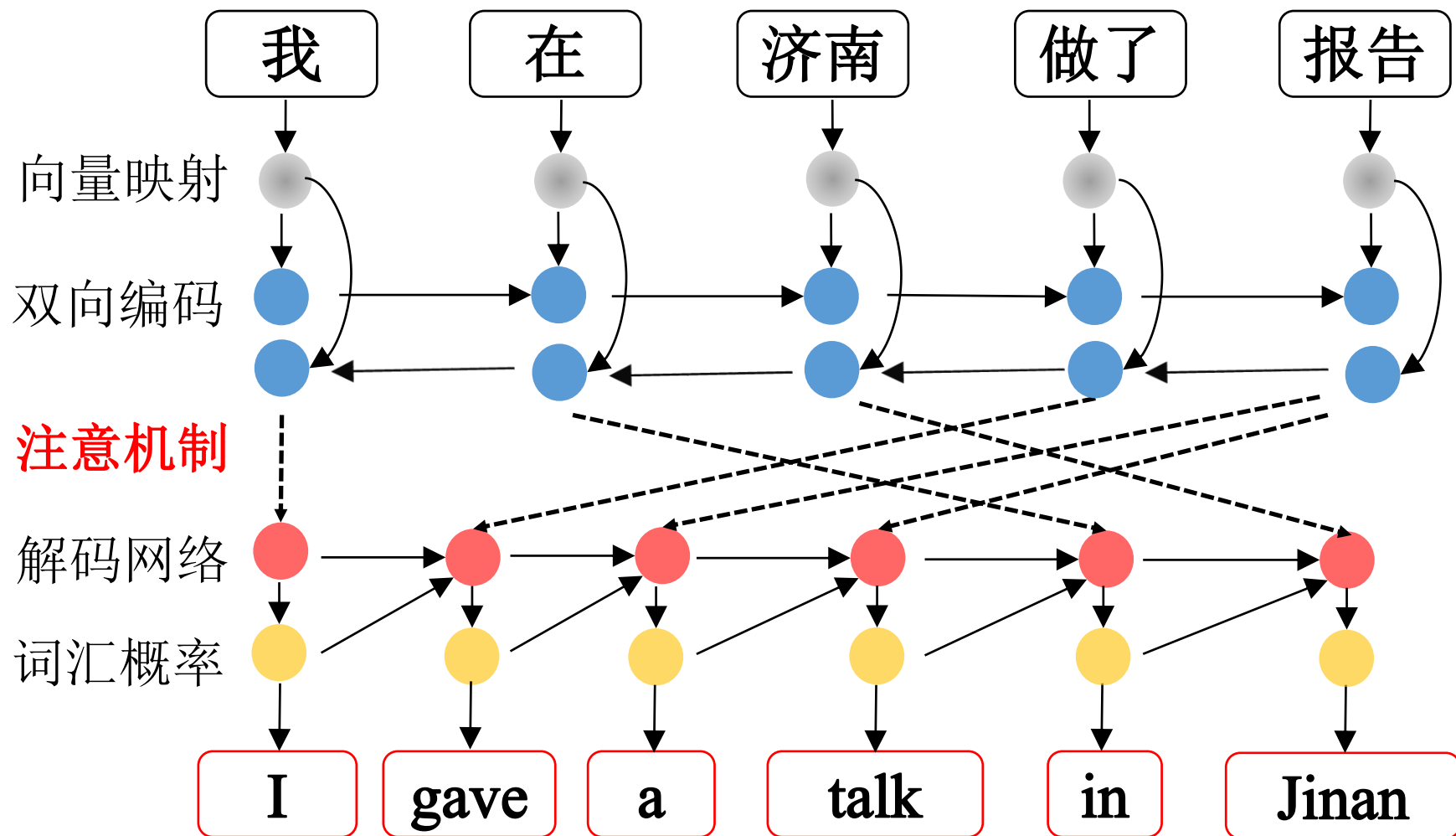
可能的目标语言句子: $E = e_1^m \equiv e_1 e_2 \cdots e_m$

首先将每个词表示成向量: Word2Vec

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

目标函数:
$$L = \sum_i \log(P(e_i | C))$$

3. 应用举例

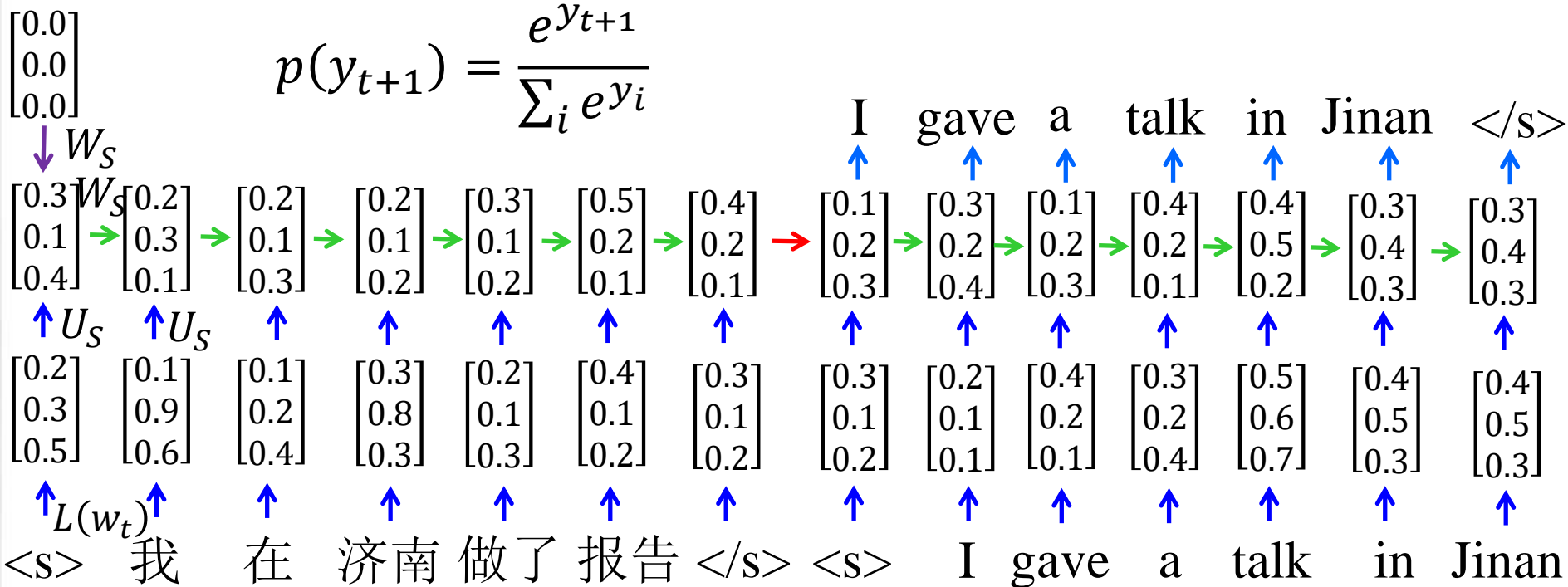


3. 应用举例

$$h_t = \tanh(U_S L(w_t) + W_S h_{t-1}) \quad h_t = \tanh(U_T L(w_t) + W_T h_{t-1})$$

$$y_{t+1} = L(w_{t+1}) \cdot h_t$$

$$p(y_{t+1}) = \frac{e^{y_{t+1}}}{\sum_i e^{y_i}}$$



3. 应用举例

➤ 机器翻译译文评估

- 主观评测：(1)流畅度；(2)充分性；(3) 语义保持性。
- 客观评测 (自动评价)
 - 句子错误率
 - 单词错误率 (mWER)
 - 位置无关的单词错误率 (mPER)
 - METEOR: 综合考虑译文词汇的准确率、召回率、F值等
 - BLEU评价指标
 - NIST 评价指标
 -

3. 应用举例

- **BLEU**评价方法 [Papineni et al., 2002]

- **Bi**Lingual **E**valuation **U**nderstudy

基本思想:

将机器翻译产生的候选译文与人翻译的多个参考译文相比较，越接近，候选译文的正确率越高。

实现方法:

统计同时出现在系统译文和参考译文中的 n 元词的个数，最后把匹配到的 n 元词的数目除以系统译文的单词数目，得到评测结果。

[Papineni et al., 2002] K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proc. ACL'2002*, pp. 311-318

3. 应用举例

如，源文：席子上有一只猫。

系统译文：the cat met a mouse in street

参考译文1：The cat is on the mat

参考译文2：There is a cat on the mat

系统译文中共有7个词，如果 n 取1的话，有3个译文词出现在参考译文中，那么， $BLEU = 3/7 = 0.43$ 。

极端情况下，系统译文为：the the the the the the the
那么，该候选译文的打分为：7/7。显然，这种译文毫无意义，打分不合理。

随着 n 值的增大，BLEU 值几乎成指数级下降，因此，BLEU方法中采用了修正的 n 元语法精度的对数加权平均值，相当于对修正的精度值进行几何平均， n 值最大为4。

3. 应用举例

另外，句子的长度对BLEU评分也有影响，如果一个机器翻译系统只翻译最可靠的词汇，译文句子就可能比较短，按上述方法计算出的精度值就会较高。

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

长度过短句子的
惩罚因子

$$w_n = 1/N$$

最大语法的阶
数，实际取4。

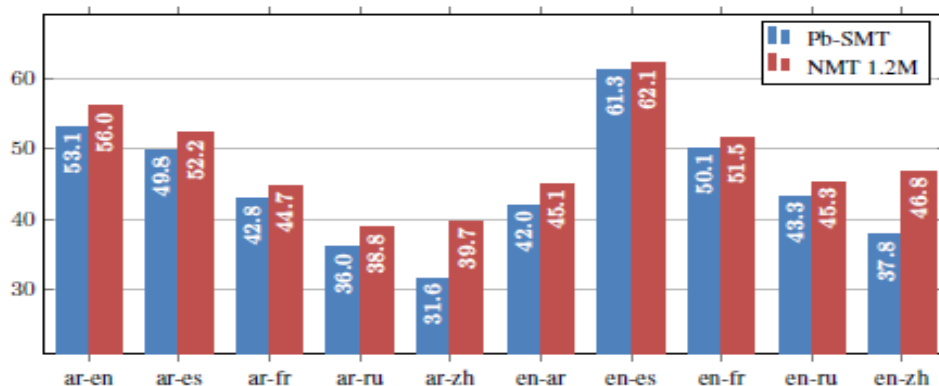
出现在答案译文中的 n
元词语接续组占候选译
文中 n 元词语接续组总
数的比例。

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

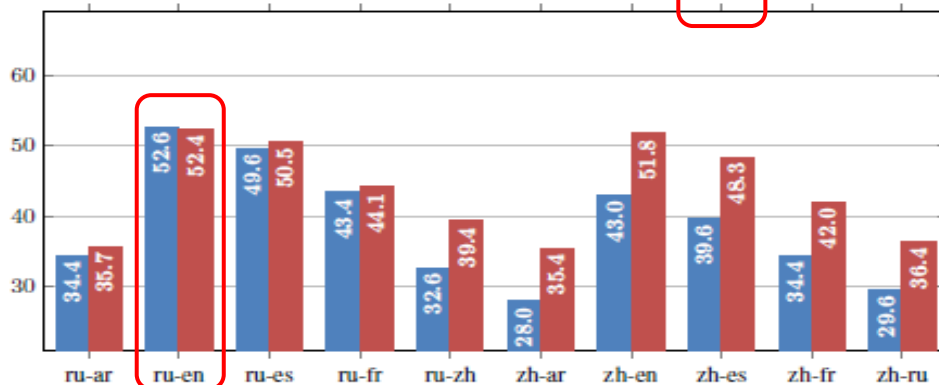
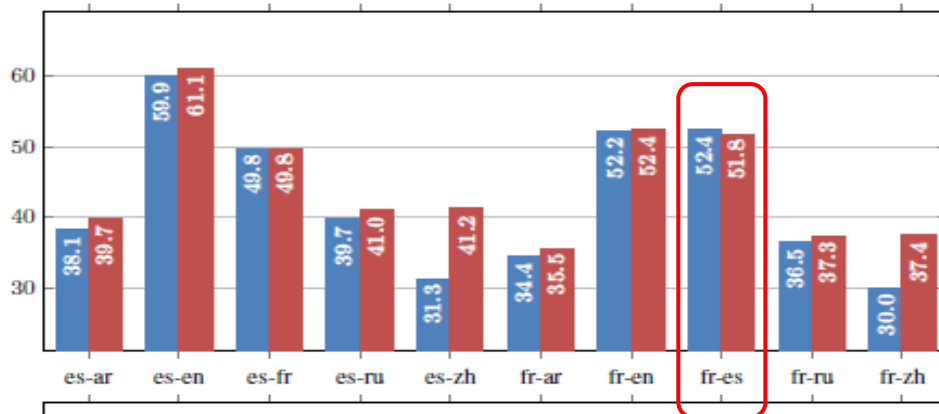
c 为系统译文中单词的个数， r 为答案
译文中与 c 最接近的译文单词个数。

BLEU 分值范围：0 ~ 1，分值越高表示译文质量越好，分值越小，译文质量越差。

3. 应用举例



神经机器翻译几乎全面超越传统的统计翻译方法。



Marcin Junczys-Dowmunt,
Tomasz Dwojak and Hieu Hoang,
2016. Is Neural Machine Translation
Ready for Deployment? A Case
Study on 30 Translation Directions.
<https://arxiv.org/pdf/1610.01108.pdf>

3. 应用举例

➤ 机器翻译译文质量

从NIST03中随机选取了500个句子进行翻译测试，83个句子出现了112处较为明显的错误。

错误类型	个数	比例
漏翻	31	27.7%
重翻	9	8.0%
错翻	21	18.8%
命名实体翻译错误	32	28.6%
调序错误	10	8.9%
语法错误	6	5.4%
成语的翻译错误	3	2.7%

“信、达、雅”是人类翻译追求的目标，目前机器翻译的译文质量基本挣扎在“信”的水平上，要做到“雅”，在可预见的未来还很难看到有较大的可能性，在相当多的翻译领域和任务上，计算机要替代人恐怕永远只是一个梦想。

3. 应用举例

- ◆ 汉语自动分词
- ◆ 机器翻译
- ◆ 问答/对话系统
- ◆ CASIA 相关工作

3. 应用举例

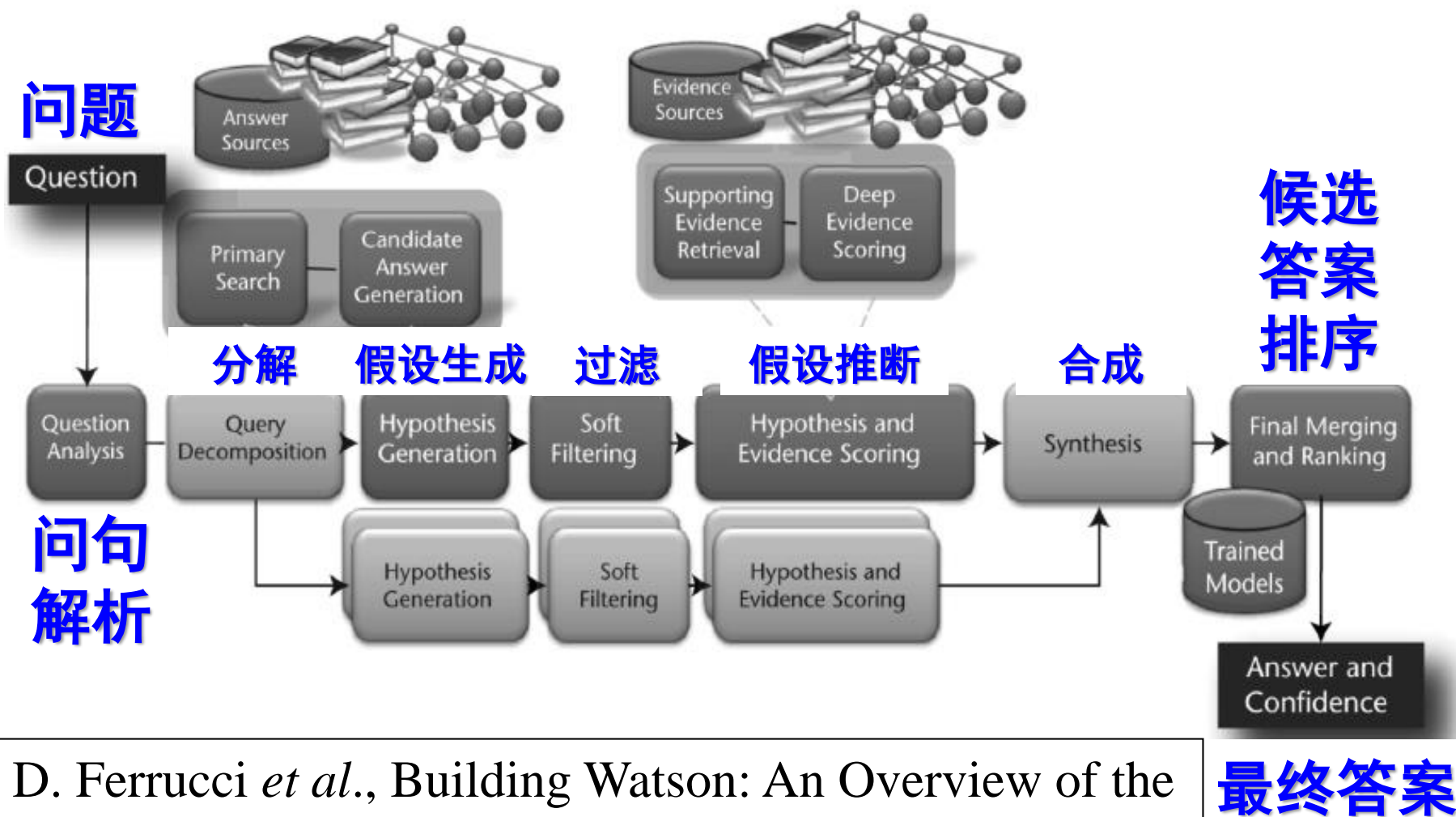
◆ 问答系统(question-answering, Q&A)



IBM“沃森”(Watson)在 2011年2月美国热门的电视智力问答节目“危险边缘”(Jeopardy!)中战胜了两位人类冠军选手。

简单的自然语言处理 + 搜索

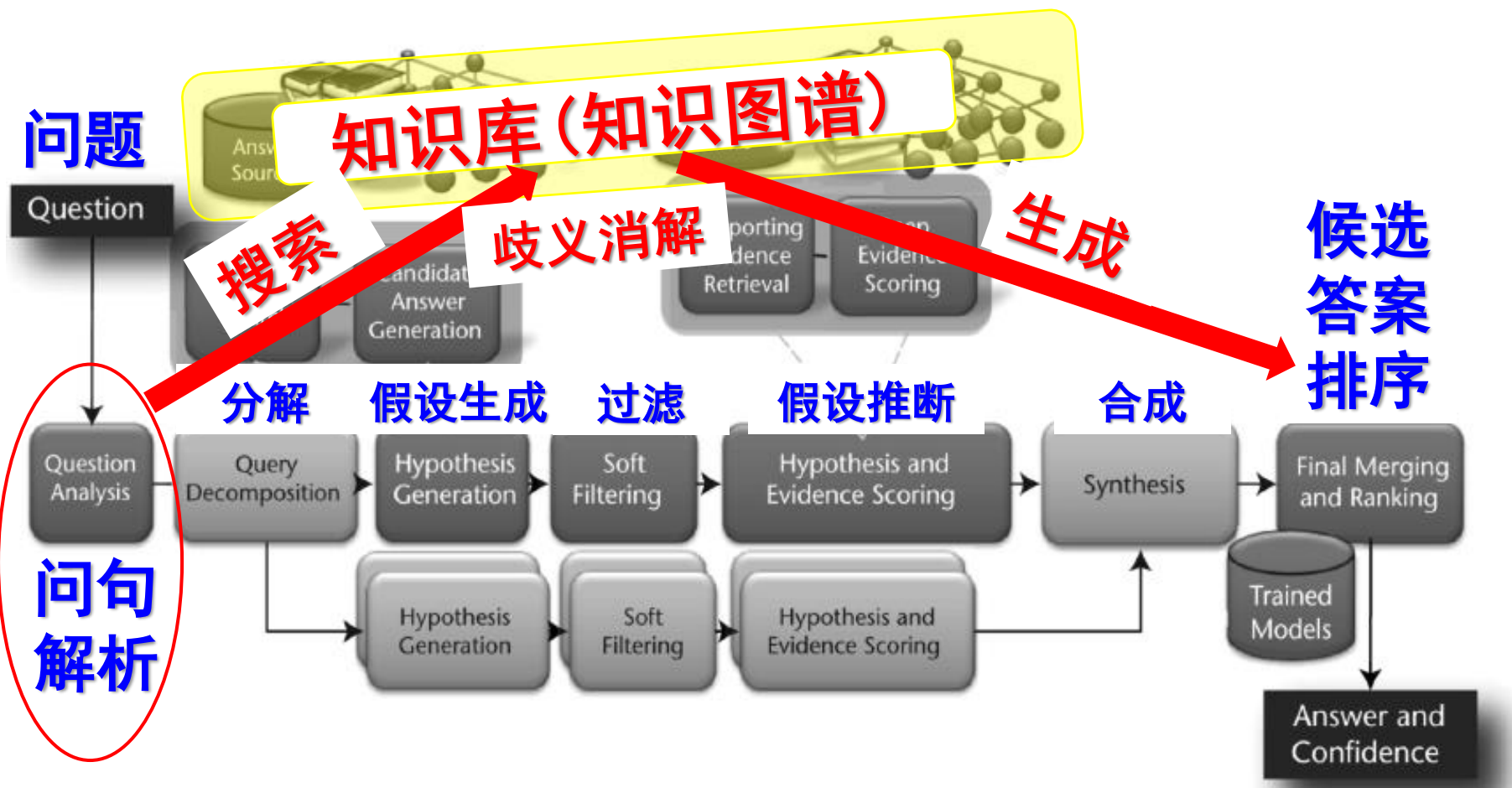
3. 应用举例



D. Ferrucci *et al.*, Building Watson: An Overview of the DeepQA Project. AI Magazine, Fall 2010, pp.59-79

最终答案

3. 应用举例

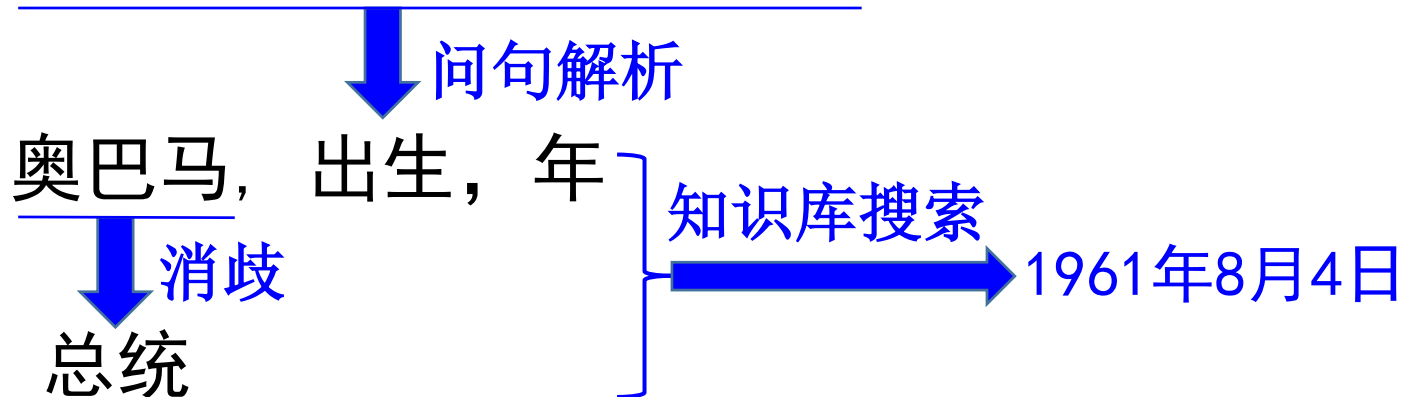


D. Ferrucci *et al.*, Building Watson: An Overview of the DeepQA Project. AI Magazine, Fall 2010, pp.59-79

最终答案

3. 应用举例

问题： 奥巴马总统出生于哪一年？

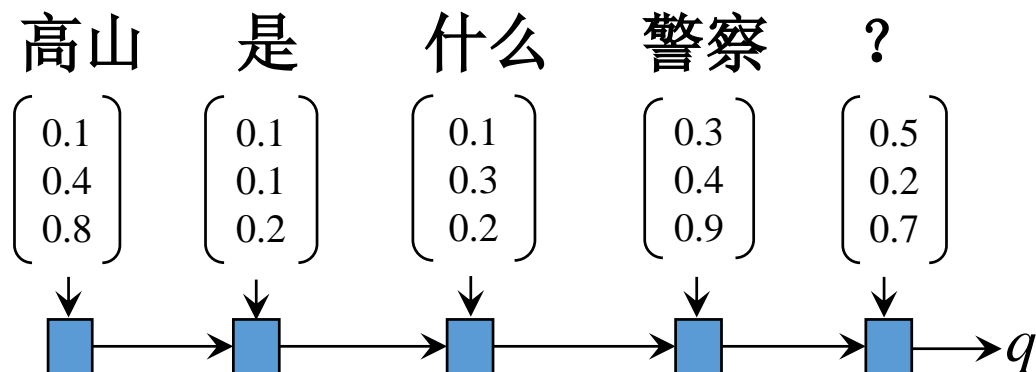


挑战： 为什么总统夫人米歇尔·奥巴马曾被媒体称为“易怒的黑人女人”？

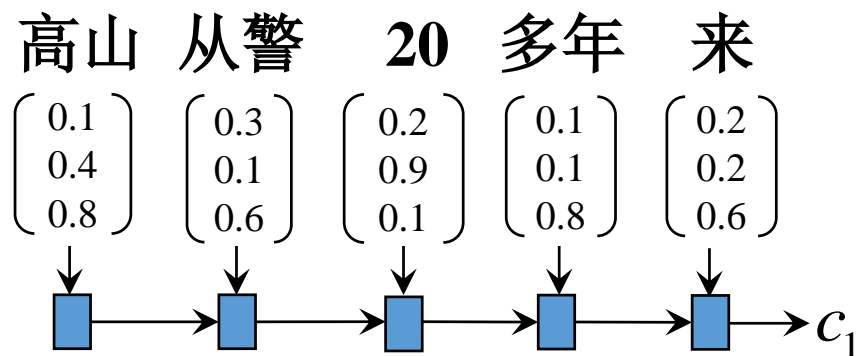
3. 应用举例

➤ 基于深度学习的问答系统原理

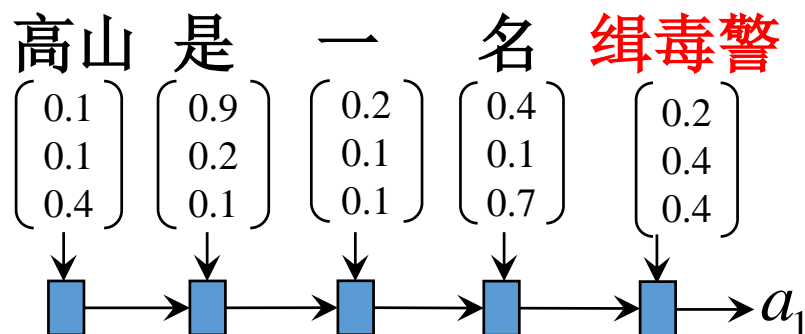
① 问句向量表示



② 上下文向量表示

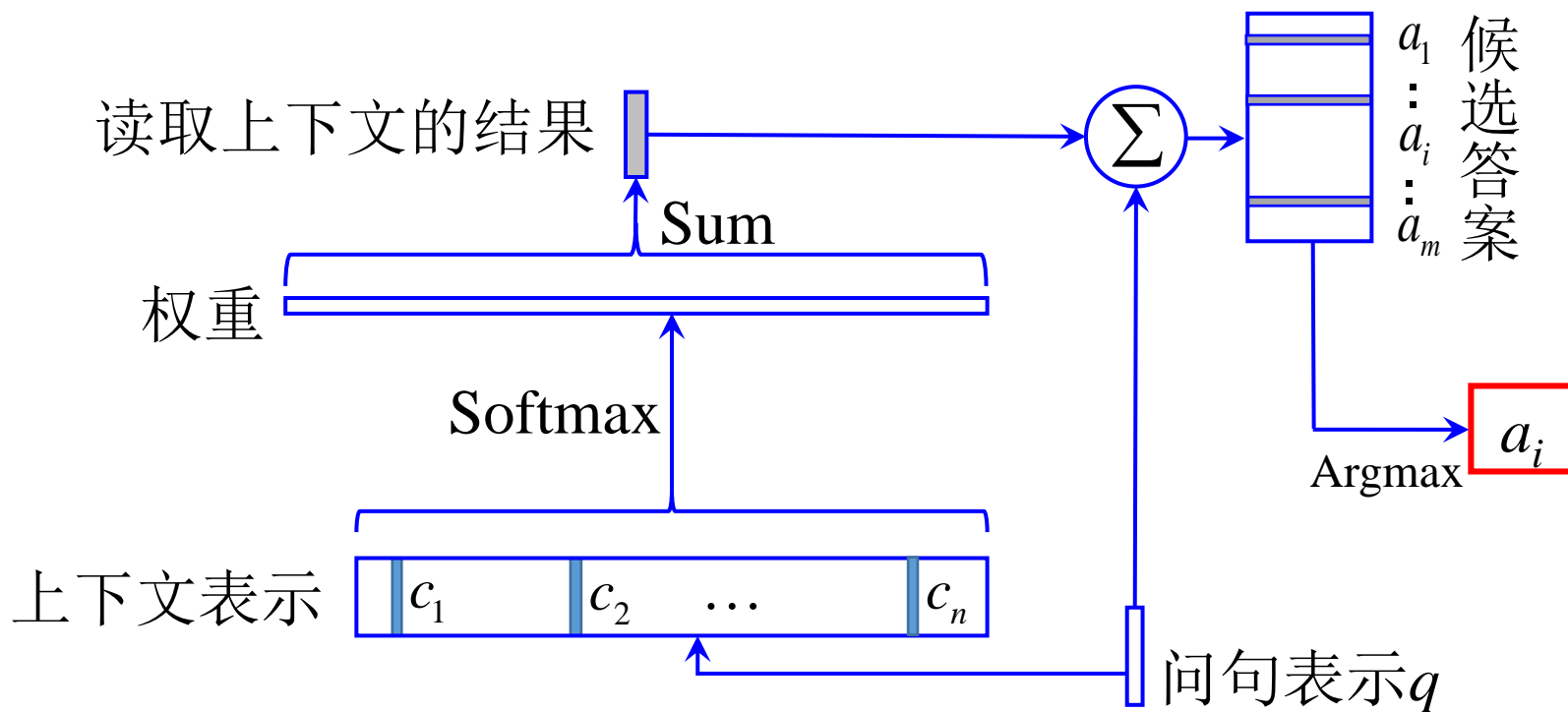


③ 候选答案表示



3. 应用举例

④ 答案推断



3. 应用举例

- ◆ 汉语自动分词
- ◆ 机器翻译
- ◆ 问答/对话系统
- ◆ **CASIA 相关工作**

3. 应用举例

◆ CASIA多语言信息翻译与分析系统

已建立15种语言、17种文字的信息翻译和分析系统，多次在国际口语翻译和全国机器翻译评测中取得第一名的优异成绩，成功应用于国家多个特定领域。



3. 应用举例



【原文】 已输入 2917 字符

[人名: 红色] [地名: 蓝色] [机构名: 绿色]

清除

【译文】

翻译

结果分析:

实体抽取

实体关系

情感分析

自动摘要

"Besides, cutting emissions has been an irreversible international trend, though there might be some temporary setbacks," Tao said.

The California government and the Ministry of Science and Technology inked an agreement on Tuesday for greater collaboration on countering climate change, according to AP.

Brown signed similar agreements in the past several days with leaders in Jiangsu and Sichuan provinces, the report said.

President Xi Jinping meets with California Governor Jerry Brown of the United States at the Great Hall of the People in Beijing on Tuesday. Xi Jinping encouraged the state to participate in the Belt and Road Initiative.

中国上周宣布，它将继续执行《巴黎气候协定》的承诺，并“逐步走向2030年目标”。

据中国社会科学院的一位研究人员介绍，中国社会科学院的研究人员说：“中国一直严肃对待它所做的承诺。”

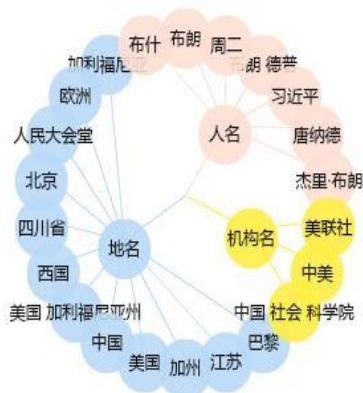
他说：“此外，减排是一个不可逆转的国际趋势，虽然可能会出现一些暂时的挫折。”

据美联社报道，加州政府和科技部周二签署了一项加强合作应对气候变化的协议。

在过去的几天里，他与江苏和四川省的领导人签署了类似的协议。

习近平星期二在北京的人民大会堂会见了美国加利福尼亚州州长杰里·布朗。习近平鼓励国家参与“一带一路”倡议。

实体抽取结果：



3. 应用举例

Egypt

Hai Yete Television Station

Ganzouri

埃及私营的**海耶特电视台**当晚报道说，**詹祖里**已被授权组建政府。前总统**穆巴拉克**执政期间，**詹祖里**于1996年1月至1999年10月任总理。

Mubarak

محمد حسنى سيد مبارك



- 1928年5月4日生于埃及曼努菲亚省米塞利赫村
- 1949年毕业于埃及军事学院
- 1950年毕业于埃及空军学院
- 1967年10月，任埃及空军学院院长
- 1969年6月任空军参谋长
- 1972年4月任埃及空军司令，同年5月兼任埃及国防部副部长



3. 应用举例



【原文】 已输入 2917 字符 [人名: 红色] [地名: 蓝色] [机构名: 绿色] 清除

"Besides, cutting emissions has been an irreversible international trend, though there might be some temporary setbacks," Tao said.

The California government and the Ministry of Science and Technology inked an agreement on Tuesday for greater collaboration on countering climate change, according to AP.

Brown signed similar agreements in the past several days with leaders in Jiangsu and Sichuan provinces, the report said.

President Xi Jinping meets with California Governor Jerry Brown of the United States at the Great Hall of the People in Beijing on Tuesday. Xi jinning encouraged the state to participate in the Belt and Road Initiative.

【译文】 翻译

中国上周宣布，它将继续执行《巴黎气候协定》的承诺，并“逐步走向2030年目标”。

据中国社会科学院的一位研究人员介绍，中国社会科学院的研究人员说：“中国一直严肃对待它所做的承诺。”

他说：“此外，减排是一个不可逆转的国际趋势，虽然可能会出现一些暂时的挫折。”

据美联社报道，加州政府和科技部周二签署了一项加强合作应对气候变化的协议。

在过去的几天里，他与江苏和四川省的领导人签署了类似的协议。

习近平星期二在北京的人民大会堂会见了美国加利福尼亚州州长杰里·布朗。习近平鼓励国家参与“带与道路”倡议。

结果分析:

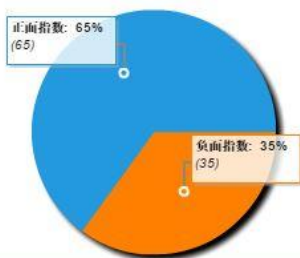
实体抽取

实体关系

情感分析

自动摘要

情感分析结果:



政治 娱乐 科技 体育 文史 教育 旅行 军事 财经

文本: 20170607-113124:President X.txt

正面指数: 65 负面指数: 35

正面示例 1: 中国上周宣布，它将继续执行《巴黎气候协定》的承诺，并“逐步走向2030年目标”。

正面示例 2: 习近平星期二与加州州长布朗(JerryBrown)进行了交谈，鼓励美国人口最多的国家为中美在关键领域的合作作出更大的贡献。

正面示例 3: 布朗先生在北京参加了一次清洁能源部长级会议，这是自从美国总统唐纳德·特朗普上周达成的《巴黎气候变化协定》以来首次举行此类会议。

负面示例 1: 他说，中国、欧洲和美国现在将填补美国联邦政府在这一问题上放弃领导权所留下的空白。

负面示例 2: 他说：“此外，减排是一个不可逆转的国际趋势，虽然可能会出现一些暂时的挫折。”

负面示例 3: 周二早些时候，布朗在第八次清洁能源部长级会议上发言时说，在第二次世界大战期间，气候变化的威胁比法西斯主义的威胁更危险。

3. 应用举例

➤ CASIA 人机交互式机器翻译系统

操作面板主界面



3. 应用举例

项目管理界面——译文审核

编辑项目[第三轮测试]

基本信息

文档管理

任务分发

成员权限

资源管理

译文审核

项目日志

机器翻译

Analysts believe the leadership will vigorously pursue reforms next year while keeping its monetary and fiscal policies broadly unchanged, so as to keep the world's second-largest economy growing within its comfort zone.

分析人士认为，明年的领导将大力推行改革其广泛的货币和财政政策不变，保持在其全球第二大经济增长的舒适地带。

句库

after the setting up of the sar next year, its legislature will be constituted and will operate according to the basic law.
香港特别行政区明年成立后，它的立法机关就按《基本法》的规定来产生和运作。
相似度:30%
新闻记忆库

the centre's provisional board, to be established next year, will need to work closely with the industrial, business and academic communities.
科技中心的临时委员会，将于明年成立，它将会与工商界及学术界保持密切联系。
相似度:29%
新闻记忆库

according to the world bank, the economy of china will become the world's second largest by 2020.
据世界银行估计，到2020年，中国经济总产量将上升到世界第二位。
相似度:30%
新闻记忆库

词典

keeping
保持; 供养; 保存; 保管; 遵守
相似度:100%
项目词典1

保管; 供养; 遵守
英汉能源大辞典

保管
英汉航海大辞典

保存; 供养; 遵守; 保持; 一致; 保管
英汉化学大辞典

供养; 遵守; 保管; 保持; 保存
计算机大词典

保管
新世纪英汉科技大辞典

学员: 所有 文档: 所有 状态: 全部

ID	原文	状态	译文
2644	China to pursue reforms while steadying growth	未锁定	中国将在保持经济平稳发展的同时追求经济改革。
2645	An annual tone-setting economic meeting, attended by top Chinese leaders, ended on Friday with a statement calling for deeper reforms while keeping growth steady and listing six major tasks for 2014.	未锁定	由中国最高领导人参加的一年一度的中央经济工作会议，在周五落下帷幕，同时发表了有关在保持经济稳定增长中深度改革的声明，并列出了2014年的六大主要任务。
2646	Analysts believe the leadership will vigorously pursue reforms next year while keeping its monetary and fiscal policies broadly unchanged, so as to keep the world's second-largest economy growing within its comfort zone.	未锁定	分析人士认为，领导人明年将在保持货币和财政政策基本不变的基础上大力推进改革，从而在中国舒适区范围内保持其在世界上第二大经济增长的地位。
2647	The statement gave no specific economic targets for 2014, while saying the government will keep the consumer price index (CPI) at around 3.5 percent.	未锁定	声明中没有指出2014年明确的经济目标，这些目标通常在三月份公布。
2648	Analysts, however, have forecast that the Chinese government is most likely to set the same targets for 2014 as in 2013 on key indicators -- aiming for GDP growth of around 7.5 percent and capping the consumer price index (CPI), a main gauge of inflation, at around 3.5 percent.	未锁定	然而分析人士曾预测，中国政府很可能在2014年的主要指标上同2013年设置同样的目标——GDP增长约7.5%和限制消费物价指数(CPI)，即通货膨胀的一种主要计量器，在3.5%左右。
2649	Reform should be integrated into all sectors of China's economic and social development in 2014, and China should maintain continuous and stable macroeconomic policies, said the statement issued after the four-day Central Economic Work Conference.	未锁定	为期四天的中央经济工作会议结束后有声明发表说，2014年改革应纳入中国经济和社会发展的所有领域中，并且中国应保持持续稳定的宏观经济政策。
2650	To soundly manage economic work next year, "the core is to seek steady progress and promote reforms and innovations," it said.	未锁定	声明指出，做好明年的经济工作，“核心是寻求稳定的发展，促进改革和创新”。
2651	It is necessary to fully understand the relationship between "sustainable and healthy development" and "economic output growth", and China should improve the "quality of growth" while avoiding "side-effects."	未锁定	充分理解“可持续发展”和“经济增长”的关系是很有必要的，中国应完善“增长的质量”，同时避免“副作用。”
2652	Also, China should continue implementing proactive fiscal and prudent monetary policies in 2014, added the statement.	未锁定	此外，中国在2014年应继续实施积极的财政政策和稳健的货币政策，声明中补充说。
	It also outlined six major economic tasks for 2014 -- ensuring food security, reducing overcapacity, containing debt		声明中还列出了2014年的六项主要经济任务——确保粮食安全、减少过剩产能、抑制债务风险、提高区域发展和民生福利以及保障

第1/623页 共6225条记录

译文审核


段落: 1639

未锁定

已通过审核

该句子可能已经审核通过，不能编辑。

内容提要

1. 问题的提出
2. 自然语言处理方法
3. 应用举例
-  4. 技术现状
5. 结束语

4. 技术现状

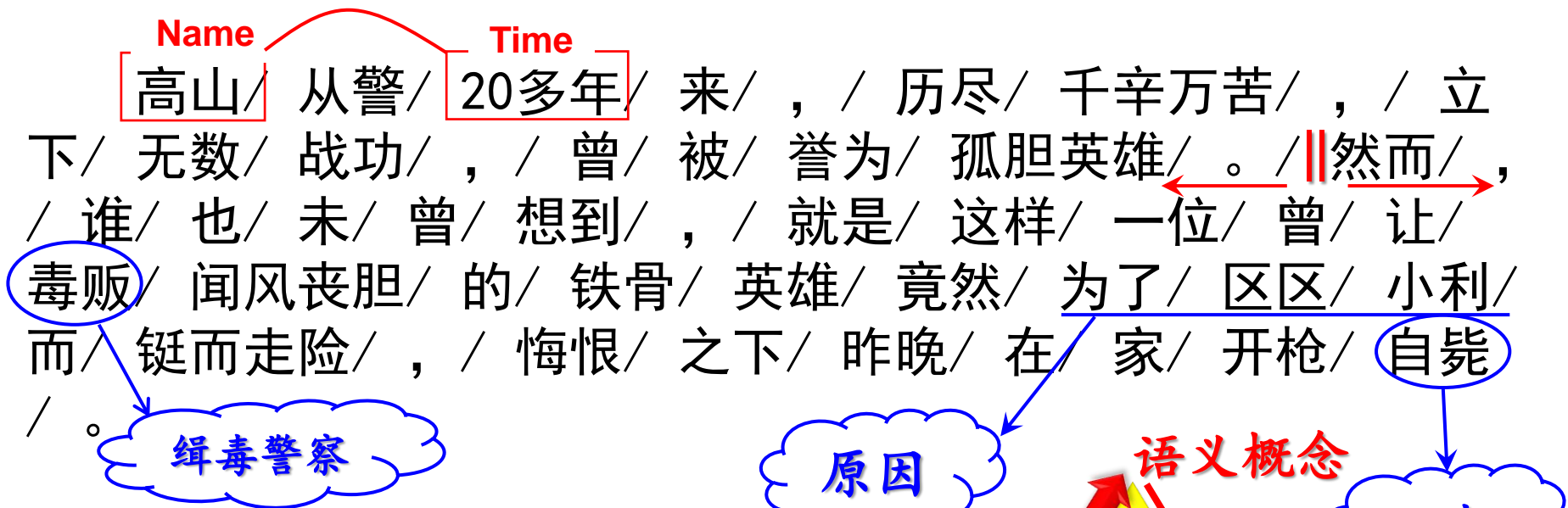
◆ 远未做到语言的深度理解

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。

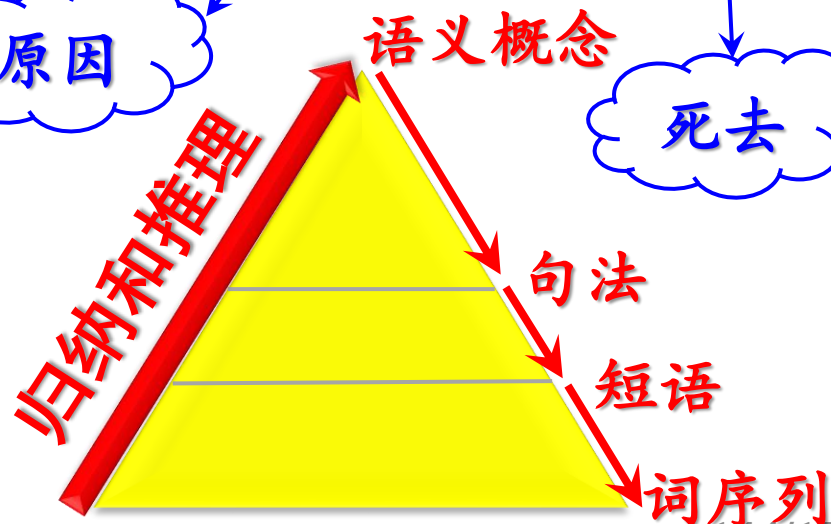
- 高山是什么警察？
- 高山死了没有？
- 高山为什么自杀？

4. 技术现状

◆ 远未做到语言的深度理解



- 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%)

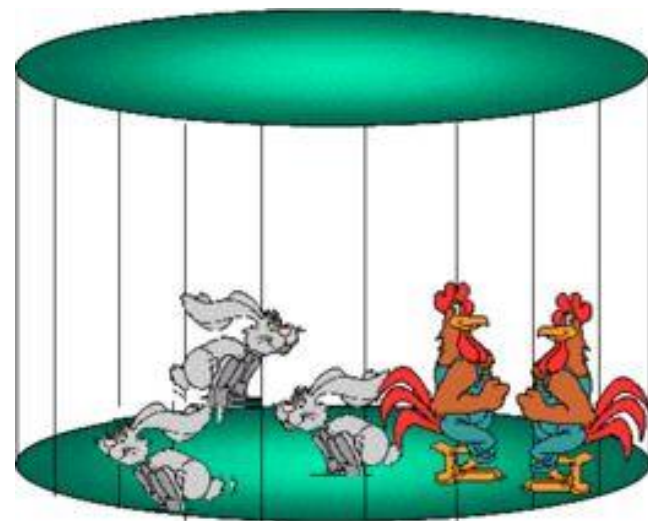


4. 技术现状

◆ 缺乏基本的常识学习能力

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




4. 技术现状

◆ 基本现状

- 部分问题得到了解决，可以为人们提供辅助性帮助，如：专业领域文档翻译，电子词典，搜索引擎，文字录入等；
- 基础问题研究仍任重而道远，如：语义表示和计算、高质量的自动翻译等；
- 社会需求日益迫切：信息服务、通讯、网络内容管理、情报处理、国家安全等；
- 许多技术离真正实用的目标还有相当的距离，尚未建立起有效、完善的理论体系。

内容提要

1. 问题的提出
2. 自然语言处理方法
3. 应用举例
4. 技术现状
-  5. 结束语

5. 结束语

- ◆ 自然语言处理是个朝阳学科，任何人都可以找到自己感兴趣的内容，无论是从事基础研究，还是应用技术开发
- ◆ 网络和移动通信技术为自然语言处理提供了巨大的潜力，从普通用户的个性化信息服务，到打击网络犯罪，维护国家安全，只要有人说话的地方，就需要自然语言处理技术
- ◆ 计算机在很多方面早就超过了人类，如存储、计算、搜索等，但在自然语言理解方面目前说已经超过了人类，恐怕只是一种幻想
- ◆ 目前无论多么强大的理论方法，包括深度学习，在自然语言处理中都显露出其无能的一面，这为自然语言理解理论研究和技术开发都留下了足够大的拓展空间

5. 结束语

- ◆ 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
- ◆ 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



N L P R



谢谢!
Thanks!