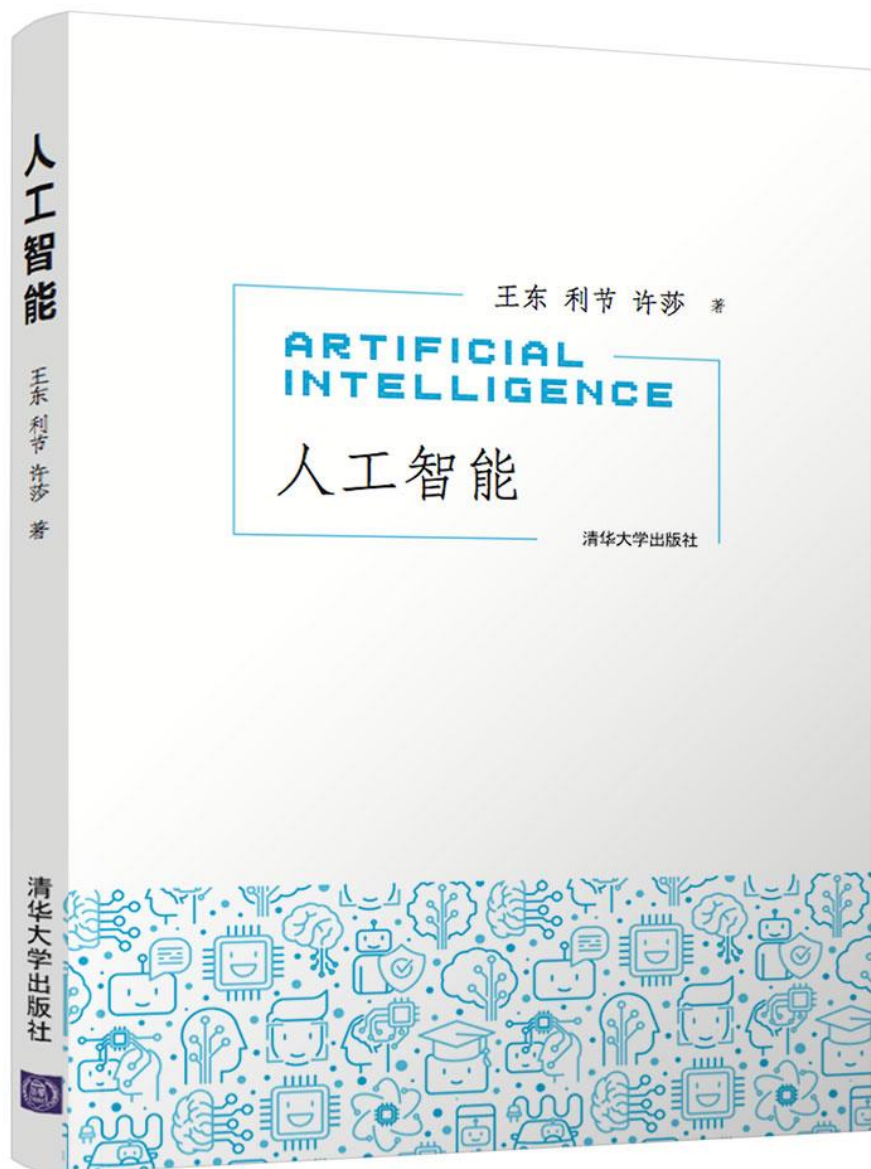


人工智能

王东 & 利节 & 许莎



- 教材名称：人工智能
- 出版社：清华大学出版社
- 作者：王东 利节 许莎

本课程应具备的能力：

- Linux下的基本操作
- python语言基础知识

资源下载：

<http://aibook.csit.org/>



理解你的语言

李蓝天

授课提示

1. 本课件90页，建议授课时间180分钟，每页平均2分钟。重点可放在“人类语言的复杂性”和“基于深度学习的语言理解方法”两部分，以提高学生的学习兴趣。
2. 若授课时间为90分钟，请去掉非课本内容，对“附注”中所提内容尽量简化，并对“传统语言理解方法”中的技术性内容进行缩减。
3. 课件内的图片、视频等素材具有较好的展示效果，建议保留。
4. 建议授课过程中对每个章节末尾的思考问题与学生进行讨论，激发学生思考。

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传 统 语 言 理 解 方 法
- 基 于 深 度 学 习 的 语 言 理 解 方 法
- 机 器 翻 译
- 语 言 理 解 的 其 他 应 用

目 录

- 引言
- 人类语言的复杂性
- 传统语言理解方法
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用

人类智能 — 语言

- 语言是人类最重要的交际工具，是人类进行沟通的主要表达方式。
- 语言伴随着人类社会而产生和发展。



语言的起源

- 起源于20万年至12万年前的非洲现在智人和尼安德特人。



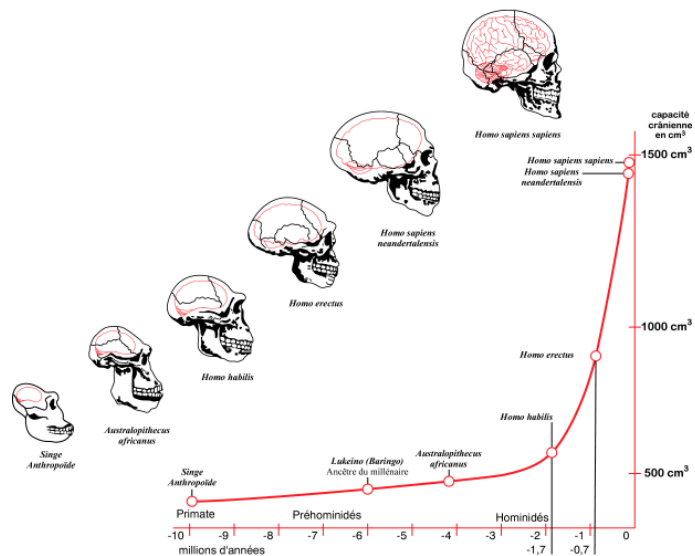
智人（拉丁学名：**Homo sapiens**），生物学分类中人属中的一个“种”，为地球上现今人类的祖先。



尼安德特人（拉丁学名：**Homo neanderthalensis**），现代人类的基因有1%~4%是来源于尼安德特人，特别是欧洲人后裔。

语言的进化

- 人类生理机能的进化
 - 大脑（复杂、高级）
 - 生理器官（独有的发声系统）

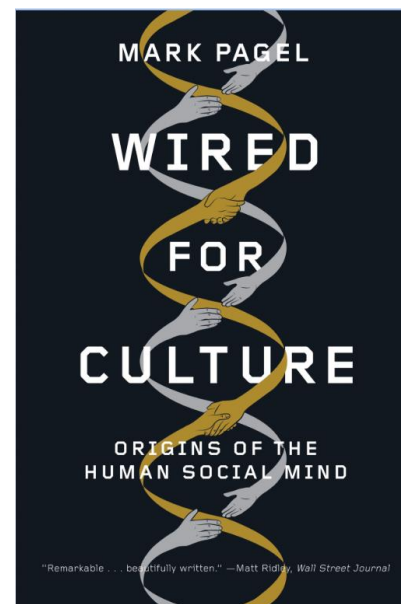


- 人类行为能力的进化
 - 符号象征意义的运用
 - 复杂的社会行为（交流、协作）



人类的“母语”

- 美国加州斯坦福大学的梅瑞特·鲁莱恩发现，tok、tik、dik和tak 反复出现在各种不同的语言中，代表了手指、脚趾、数字等。
- “妈妈”是人类语言中最早出现的名词，世界几乎所有语言中都有“m”这个音节。
- “huh”（哼、哈）全球通用词语，也是最早出现的人类词语之一。



Wired for Culture: Origins of the Human Social Mind
Mark Pagel

复杂多样的语言

- 7000 多种语言
- 2000 多种语言有书面文字
- 发音独特、规则各异
- 人类与生俱来的语言学习能力



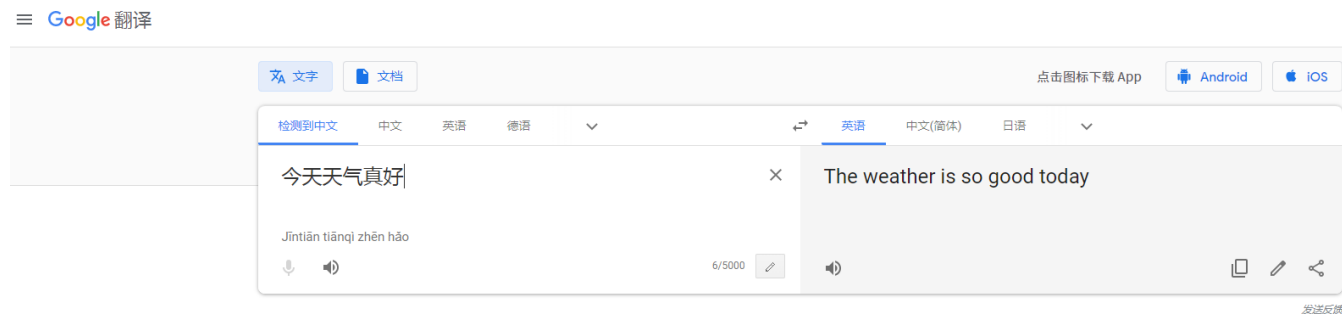
机器如何理解语言？



Baidu 搜索引擎



IBM 沃森 DeepQA 问答系统



Google 翻译

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传 统 语 言 理 解 方 法
- 基 于 深 度 学 习 的 语 言 理 解 方 法
- 机 器 翻 译
- 语 言 理 解 的 其 他 应 用

A. 结构复杂性

我们认真地努力地工作，就会做出更大贡献，让我们的国家更强大。否则，别的国家就会看不起我们。

- 丰富的层次结构和复杂的依赖关系
 - 第一层：一个段落包含若干个以语义关系组织起来的句子，组成表达单元
 - 第二层：每个句子包含若干小句，组成完整的意思表达
 - 第三层：每个小句由若干单词按语法结构组合而成
 - 第四层：每个词由汉字按一定成词规则组合而成
 - 第五层：基本词素单元

A. 结构复杂性

■ 语言表达的层次性

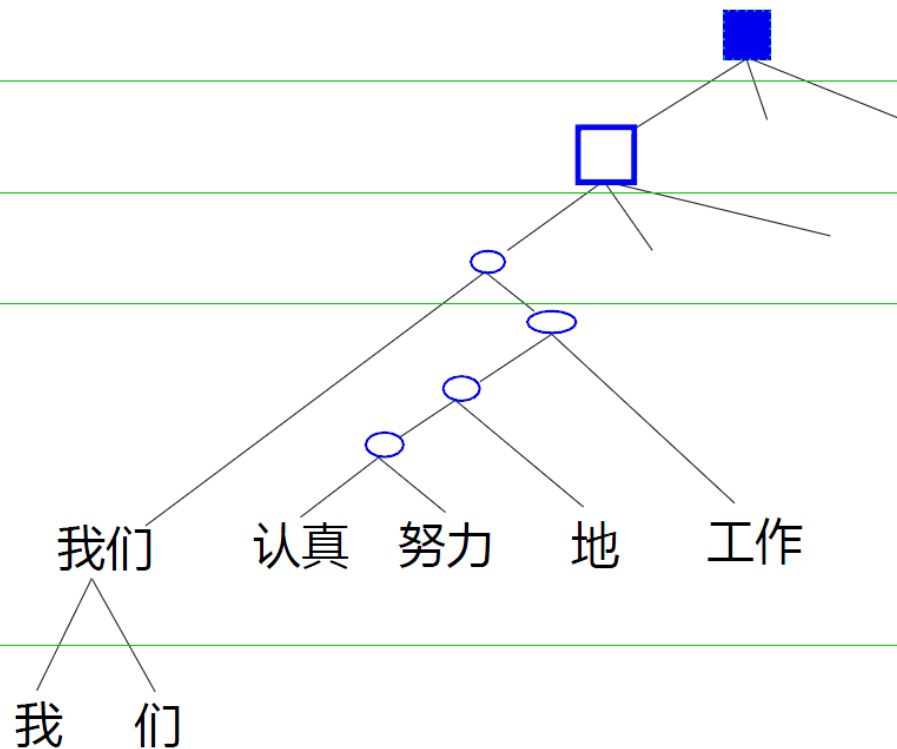
段落

句子

小句

词

汉字



形简意丰：

形式上是一串文字，但蕴含了复杂的语义和逻辑内容

B. 语义复杂性

- 词本身的歧义性和词与词的组合结构歧义性

我想起来了 “我/想起来/了” “我想/起来了” (“想” 和 “起来”) 的歧义性

他在那儿看东西 “看守东西” “用眼睛看东西” (“看东西”) 的歧义性

我要榨果汁 “我想要自己榨杯果汁” “我想点一杯榨果汁” (“榨果汁”) 的歧义性

研究所有东西 “研究所/有东西” “研究/所有/东西” (“研究所有”) 的歧义性

C. 知识复杂性

- 语言中所包含的知识

煤中含碳元素，不充分燃烧时会产生一氧化碳。一氧化碳进入人体和血红细胞结合，使血红细胞失去携氧功能，产生一氧化碳中毒。因此，我们在烧煤时要尽量保持通风，特别是在低压天气，要特别注意。

- 缺少专业知识，使得理解变得困难！

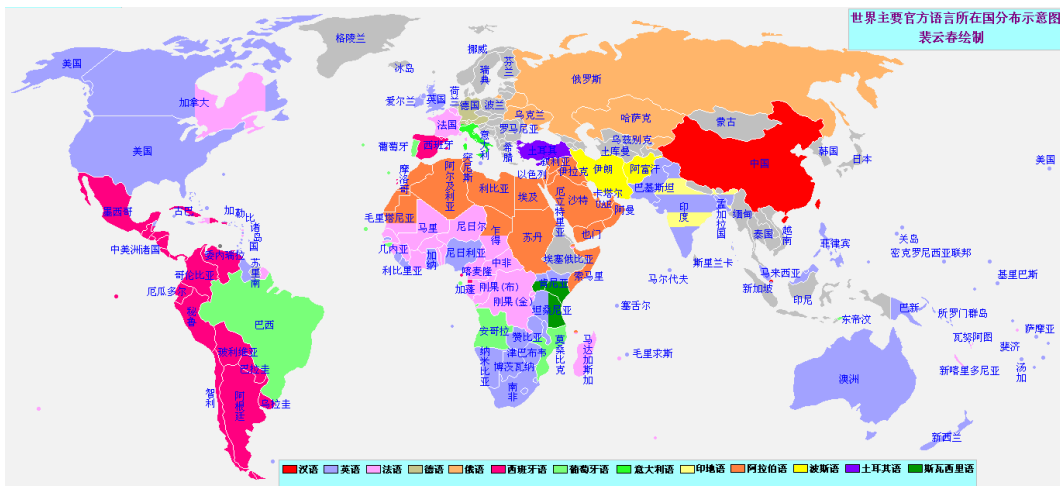
D. 时空复杂性

■ 不同地域

- 大陆中文 vs 台湾中文
- 东北话 vs 河南话
- 美国英语 vs 印度英语

■ 不同时代

- 古代汉语 vs 现在汉语
-> “北出牖也，从宀从口”
- 现代汉语 vs 网络俚语
-> 微我、么么哒



E. 应用复杂性

■ 语言的不规范性

GG: 你嚎

MM: 你嚎，你在哪里伤亡？

GG: 在王八里上网，你呢？

MM: 也在王八里

GG: 你是哪里人

MM: 是鬼洲人

GG: 哦，是山洞人

MM: 你似男似女？

GG: 当然是难生了，你肯定是驴身了

MM: 是

GG: 你霉不霉？

MM: 还行吧，人家都说是大霉女，你衰不衰？

GG: 还好，很多人都说是大衰哥

MM: 真的呀，咱们多怜惜好不好？

GG: 好鸭，你的瘦鸡号码是多少？

MM: 咱别用瘦鸡，瘦鸡聊天多贵，就用你扣扣

GG: 你真可爱，很想煎你

MM: 慢慢来，虽然隔得远，蛋也有鸡会。

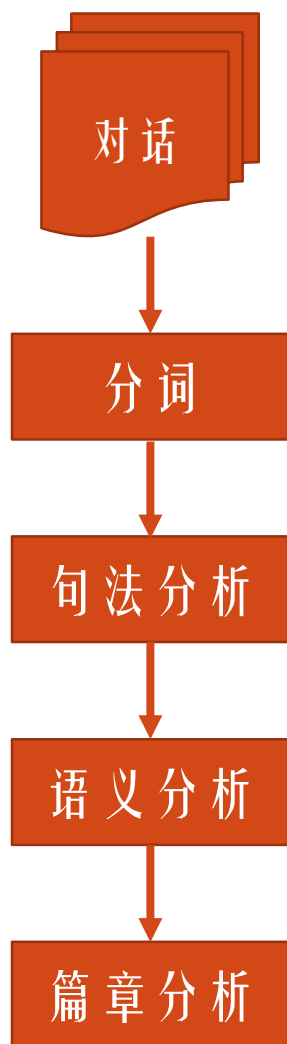
什么是语言理解？

- 计算机 + 语言学
- 目标：使得计算机能够“处理”并“理解”自然语言，从而帮助人类完成一些任务。
- Perfect language understanding is AI-complete.

机器如何理解语言？

- a. 解析思路

- 细致拆解，直到细节（刨根问到底）
- 传统自然语言理解方法



A: 北京真是个好地方!
B: 是吗? 为什么?
A: 因为北京很方便啊
B: 我没觉得呀!

北京 / 真是 / 个 / 好地方!

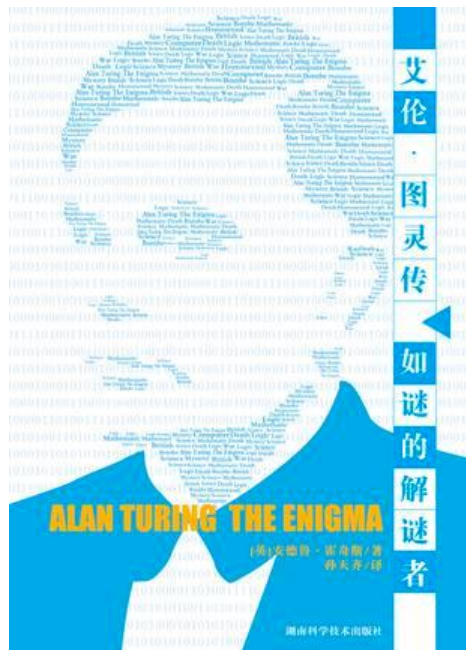
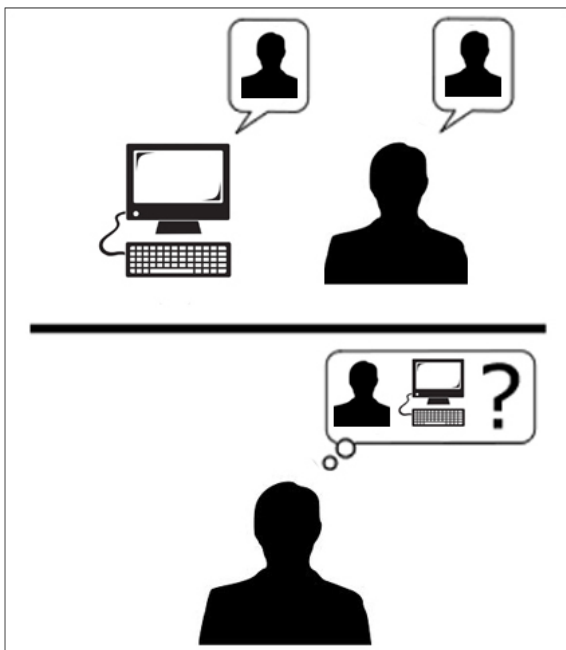
主语	谓语	宾语
北京	真是	个 / 好地方!

问: A 喜欢北京吗?

问: A 和 B 都喜欢北京吗?

机器如何理解语言？

- b. 反馈思路
 - 不问方式，不纠细节，只看结果
 - 图灵测试：测试机器是否具备人类智能



```
Welcome to

EEEEEE LL      IIII ZZZZZZZ AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII  ZZZZZZZ AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

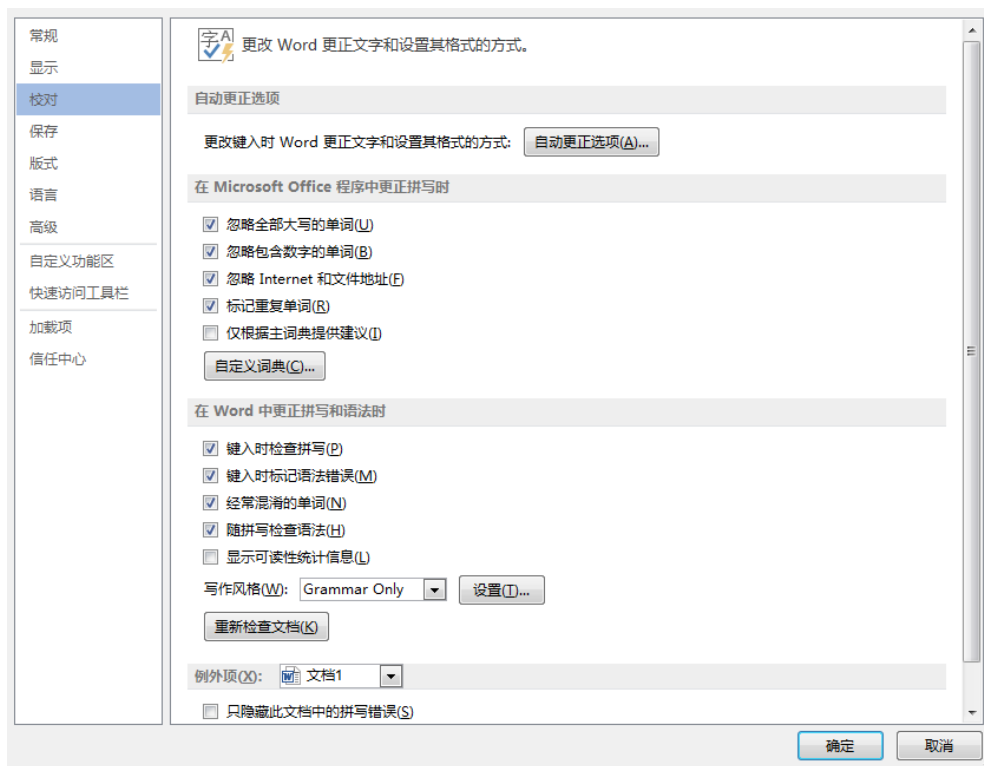
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

1964-1966年的人机会话程序
ELIZA 是基于反馈思路设计的

<http://www.masswerk.at/elizabot/>

语言理解的应用：一

■ Word 自动检查拼写错误



今天天气好真呀

今天天气真好呀

语言理解的应用：二

■ 搜索引擎：Google、Baidu...

Google

中国

全部 地图 图片 新闻 视频 更多 设置 工具

找到约 4,380,000,000 条结果 (用时 0.48 秒)

焦点新闻

武汉新型冠状病毒：中国确诊病例激增，习近平要求遏制疫情，香港监控范围扩大至湖北
BBC.com
2 小时前

肺炎疫情 中国醒了 习近平发话后李克强开腔 首席专家急赴武汉
RFI
1 小时前

中国出生率创历史新低，人口危机继续逼近
纽约时报
11 小时前

→ 更多关于“中国”的结果

baike.baidu.com · item · 中国

中国（世界四大文明古国之一）_百度百科

中国文化源远流长、博大精深、绚烂多彩，是东亚文化圈的文化宗主国，在世界文化体系内占有重要地位，由于各地的地理位置、自然条件的差异，人文、经济方面也各有 ...

zh.wikipedia.org · zh-hans · 中国

中國- 维基百科，自由的百科全书

中国是位於東亞的國家或地理區域，此名稱最早見於西漢，用來指以洛陽盆地為中心的中原地區，與四夷相對的領域概念，之後因為文化的擴張，逐漸用來指稱從夏朝 ...

历史 · 文化 · 科學及技術 · 日常生活

中华人民共和国

东亚的国家

中华人民共和国，简称中国，是位于东亚的社会主义国家，首都位于北京。中国领土陆地面积约960万平方千米，是世界上陆地面积第二大、陆地面积第三大、总面积第三大或第四大的国家，其分为23个省份、5个自治区、4个直辖市和2个特别行政区。 维基百科

人口：13.86 亿 (2017 年) 世界银行， 热点

首都：北京市

货币：人民币

主席：习近平

官方语言：现代标准汉语

用户还搜索了

还有10+项

俄罗斯 伊朗 美国 香港 印度

Baidu 百度

中国

百度一下

网页 资讯 视频 图片 知道 文库 贴吧 采购 地图 更多»

百度为您找到相关结果约100,000,000个

搜索工具

中华人民共和国_百度百科

中华人民共和国（People's Republic of China），简称“中国”，成立于1949年10月1日，位于亚洲东部，太平洋西岸，是工人阶级领导的、以工农联盟为基础的人民民主专政的社会主义国家。 历史沿革 自然环境 自然资源 行政区划 国家象征 更多>> https://baike.baidu.com/

中国卫生人才网

习近平专访世卫总干事 突显中国对全球... 2017-07-19 人才杂志 更多>> 卫生健康人才发展70年 2019-09-26 探索老年健康服务体系建设 2019-09-26 对发展医... www.21wecan.com/index_... - 百度快照

中国政府网_中央人民政府门户网站

中国政府网由国务院办公厅主办，中国政府网运行中心负责运行维护，是国务院和国务院各部门以及各省、自治区、直辖市人民政府在国际互联网上发布政府信息和提供在线服务的... www.gov.cn/ - 百度快照

中国(世界四大文明古国之一)_百度百科

4天前 · 中国是以华夏文明为源泉、中华文化为基础，并以汉族为主体民族的多民族国家。通用汉语、汉字，汉族与少数民族被统称为“中华民族”，又自称为炎黄子孙、龙的传人。 ... 百度百科 - 百度快照

中国吧_百度贴吧

欢迎来到中国吧参与讨论
关注用户：94万人

相关书籍

中国概况教程
肖立奇语言
文字图书

中国文化概况
文化评述图
书

中国文化英语教程
孙克勤著
游地图

美食中国
天津古籍出版社出版

中国旅游资源概况
专业基础课
和重点课程

中国地理概况
国务院侨务
办公室所著

中国旅游客源国概况
孙克勤著
游地图

中国地形
中国人口
中国现阶段
的主要问题

其他人还搜

中国人口
中国现阶段
的主要问题

中国地图
神州大地跃
然纸上

中华人民共和国行政...
现共有34个
省级行政区

中华人民共和国国家...
明确中国版
图

语言理解的应用：三

- 机器翻译：Google 翻译、有道词典...



The screenshot shows the Youdao Dictionary web interface. On the left is a sidebar with navigation options: 查词 (Search), 翻译 (Translate), 例句 (Examples), 百科 (Encyclopedia), 单词本 (Wordbook), 人工翻译 (Human Translation), 精品课 (Premium Course), 设置 (Settings), 满意度 (Satisfaction), 取词 (Get Word), and 划词 (Highlight Word). The main content area displays the entry for '中国' (China) with its pinyin [zhōng guó] and a star icon. Below this, it shows the English translation 'China' and a link to the '新汉英大辞典' (New Chinese-English Great Dictionary). The definition provided is 'China; Sino-: road of socialism with Chinese characteristics 中国特色的社会主义道路'. It also mentions the source as '以上来源于《新汉英大辞典》'. Under the '网络释义' (Network Definition) section, there are three items: 'China', 'Chinese', and 'China Telecom'. A '短语' (Phrase) section follows, listing three items: '中国 China; Chinese; China Telecom', '中国魟 Chinese stingray; Dasyatis sinensis', and '中国国情 China's national conditions; China's existing realities; China's actual'.

查词

翻译

例句

百科

单词本

人工翻译 new

精品课 new

设置

满意度

取词

划词

中国 [zhōng guó] ☆

China

新汉英大辞典

中国 [zhōng guó]

China; Sino-:
road of socialism with Chinese characteristics 中国特色的社会主义道路

以上来源于《新汉英大辞典》

网络释义 秒懂释义

> China

> Chinese

> China Telecom

短语

1. 中国 China; Chinese; China Telecom

2. 中国魟 Chinese stingray; Dasyatis sinensis

3. 中国国情 China's national conditions; China's existing realities; China's actual

语言理解的应用：四

- 对话系统：Siri、Echo、智能客服...



思考

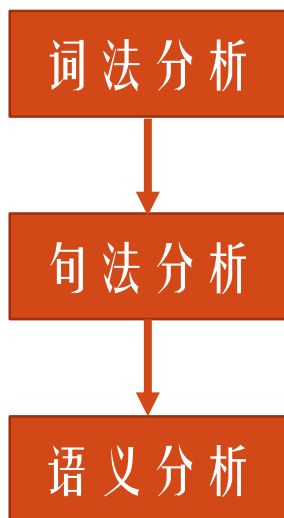
- 在生活和学习中，人类语言的复杂性还有哪些体现呢？
- 语言理解能给人类带来哪些便利或者困扰呢？

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传统语言理解方法
- 基于深度学习的语言理解方法
- 机 器 翻 译
- 语 言 理 解 的 其 他 应 用

传统语言理解方法

- 细致拆解，直到细节（刨根问到底）



北 京 / 真 是 / 个 / 好 地 方 ！

主 语	谓 语	宾 语
北 京	真 是 / 个	好 地 方 ！

问：A 喜欢北京吗？

词 是 什 么？

- 词是最小的表达单元
- 既是语义单元（**词义**）又是语法单元（**词性**）
- “走”：行走；离开；泄露（动词）
- “太阳”：物体（名词）
- “温暖”：形容词（被窝真温暖）；动词（他的言行温暖了我）

词法分析

- 词法分析是指从输入序列中确定词序列，并标记每个词的词性。

中华民族是伟大的民族



中华民族 / 是 / 伟大的 / 民族



名词



动词



形容词



名词

分词

- 不同于英语，汉语中没有明确的词边界，因此需要对汉字序列“分词”
- 分词是汉语独特的词法分析任务，具有一定的灵活性。
- 分词方法不同，语义可能也不同。

中华民族是伟大的民族

中华/民族/是/伟大/的/民族

中华民族/是/伟大的/民族

发现大道有活动

发现/大道/有/活动

发现大道/有/活动

中科院 ICTCLAS 在线分词平台

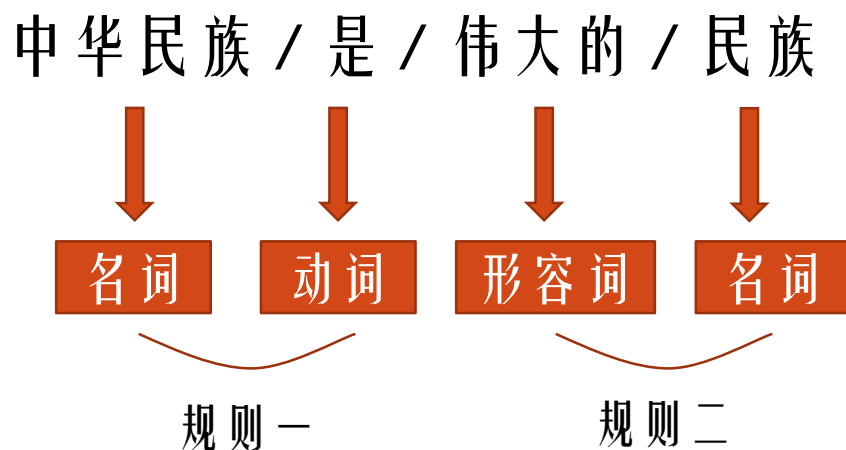
<http://ictclas.nlpir.org/nlpir/>

词性标注

- 本质是分类问题，将预料库中的词按词性分类。
- 一个词的词性由其在所属语言的含义、形态和语法功能决定。
- 词性标注与上下文相关
 - 被窝真温暖（形容词）
 - 他的言行温暖了我（动词）

标注方法

- 标注方法
 - 基于规则的方法（动词后面接名词；名词前面是形容词...）
 - 基于统计模型的方法（统计不同词性的单词所组成的最大词性序列）

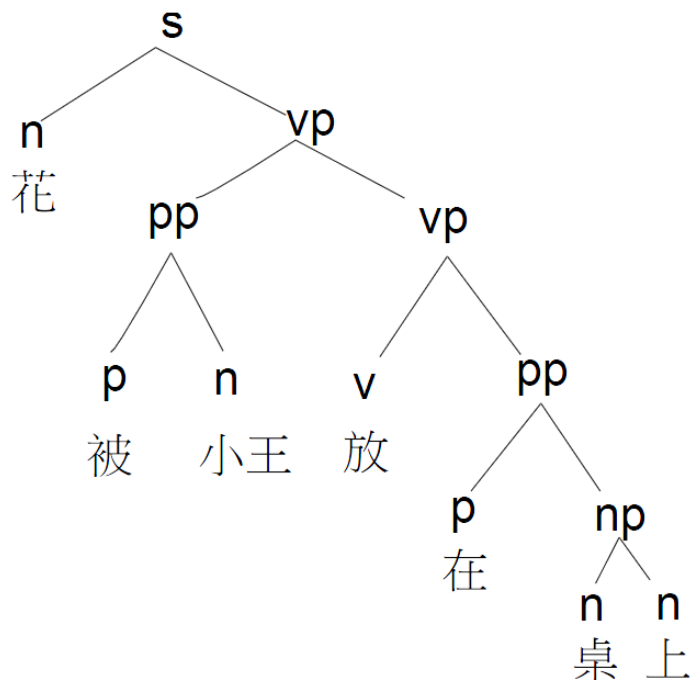


中华民族	是	伟大的	民族	概率
名词	动词	形容词	名词	0.90
形容词	动词	形容词	名词	0.08
名词	连词	形容词	名词	0.02

基于统计概率

句法分析

- 是在词法分析的基础上，对一句话中词与词的**组合方式**进行解析。
- 词与词之间通过层次性结构组合成句子，表示为一颗**句法树**。



叶子节点：词
中间节点：短语

基于规则的句法分析

- 按照某种规则生成句法树（生成句子的过程）
- 上下文无关文法（Context Free Grammar, CFG）

生成规则

逆向推理

句法分析

句法树

规则 1: $S \rightarrow N VP$;

规则 2: $VP \rightarrow V N$;

规则 3: $V \rightarrow \text{吃} \mid \text{拿}$

规则 4: $N \rightarrow \text{猴子} \mid \text{苹果} \mid \text{香蕉}$

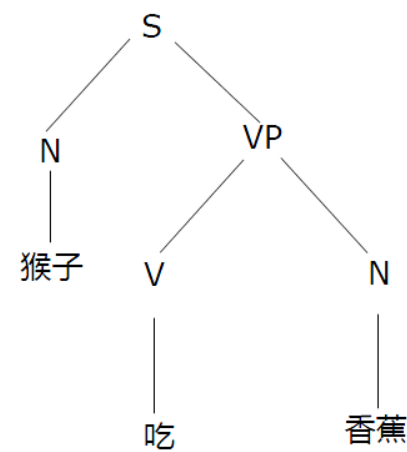
步骤 1: 猴子 $\rightarrow N$ (规则 4)

步骤 1: 吃 $\rightarrow V$ (规则 3)

步骤 2: 香蕉 $\rightarrow N$ (规则 4)

步骤 3: $V N \rightarrow VP$ (规则 2)

步骤 3: $N VP \rightarrow S$ (规则 1)



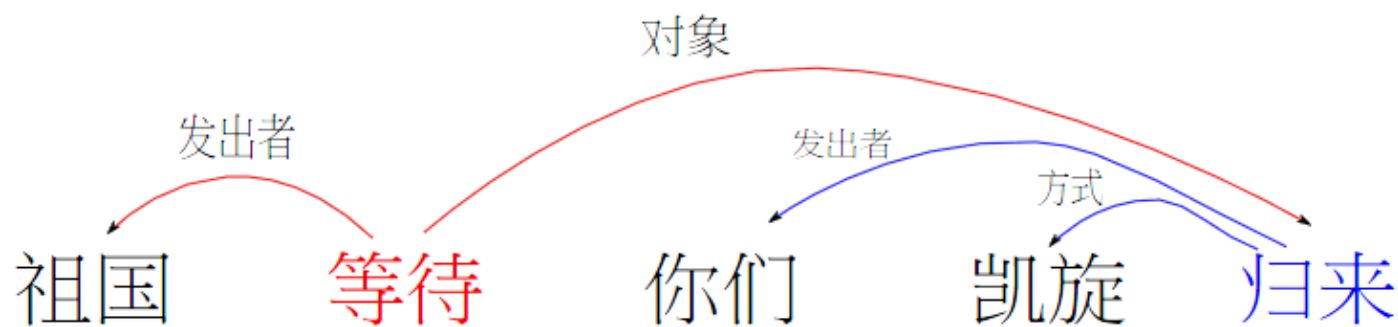
语 义 分 析

- 句法分析是对句子结构的解析，语义分析是对句子内容的解析。
- 语义分析将自然语言表达的句子转化为一种知识。

“我有一辆自行车”，陈述了一种现实，一种信息。

形式化表示

- 语义分析的代表方法
 - 格式化的、简洁的
 - 明确的、无歧义的
 - 灵活的、表达能力强的
- **语义依存树**：一句话中各个组成成分之间的语义角色关系



哈工大 ITP 语义分析平台

<http://itp.ai/demo.html>

知识图谱

- 三元组：描述两个概念之间的关系。
- 知识图谱：知识的图形化表示

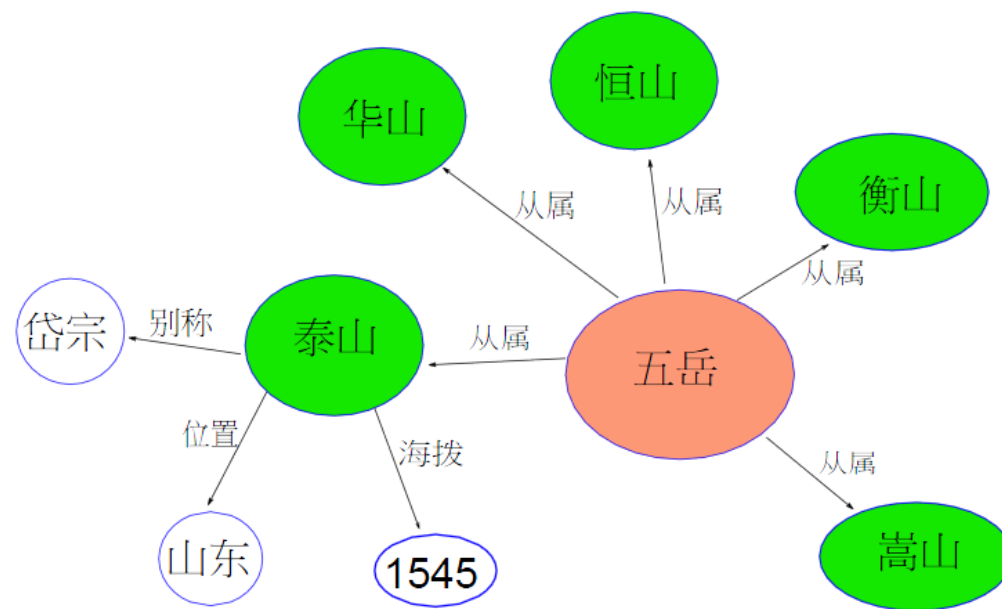
- 知识库

拥有（我，自行车）
拥有（小明，汽车）
喜欢（张亮，小云）

- 问答系统

“张亮喜欢谁”

喜欢（张亮，？）



“泰山有多高” → 海拔（泰山，1545）

思考

- 传统语言理解方法的优点和缺点？

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传 统 语 言 理 解 方 法
- 基于深度学习的语言理解方法
- 机 器 翻 译
- 语 言 理 解 的 其 他 应 用

传统 VS 深度学习

■ 传统方法

■ 优点

- 对句子深入解析，全方位理解
- 基于规则，解析偏差小
- 特定领域、简单句子效果不错

■ 缺点

- 需要丰富的领域知识
- 层次性依赖，易产生错误传导
- 自然语言变化莫测，通用性极差

■ 深度学习方法

■ 特点

- 不依赖句子解析
- 语义空间（连续、可度量）
- 语言理解的反馈思路（图灵测试）
- 不问方式，不纠细节，只看结果

传统机器与人类大脑

- 传统语言理解中的“词”

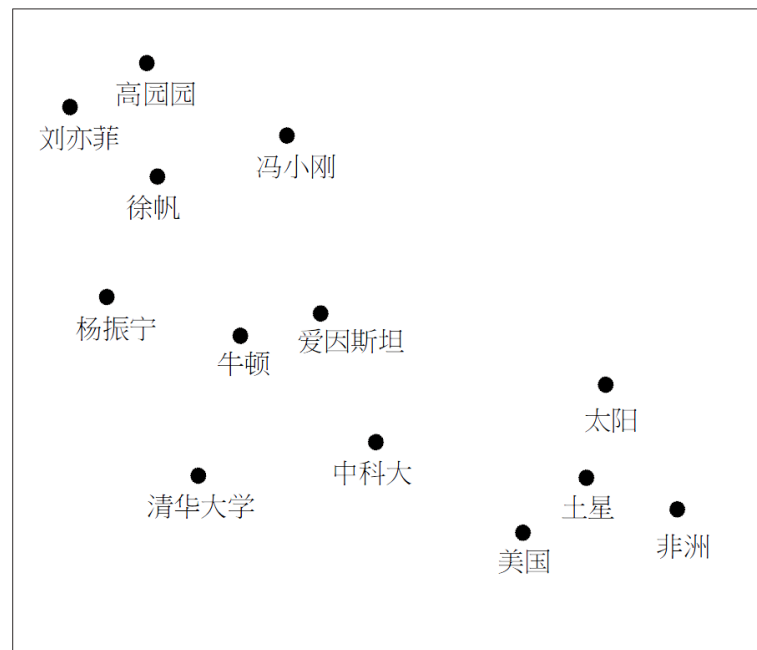
- 离散的、独立的
- 明确的属性（词性和词义）
- 难以实现知识推理

已知：“玫瑰 / 有 / 香味”

问：“杜鹃 / 有 / 香味 / 吗？”

- 人类大脑对“词”的认知

- 连续的、可度量的、可推理的
- “心理词典” -> “单词地图”



“单词地图”

语义相似，距离较近
语义迥异，距离较远

词 向 量 概 念

- 本质：是机器所绘制的人脑中的“单词地图”！
- 目标：将单词表示为一个高维空间中的连续向量，使得在该空间中语义越近的单词对应的向量距离更近。
- 意义：原本离散的单词集合变成了一个连续可度量的语义空间，使得语义的计算不再是逻辑运算，而是代数运算！

词 向 量 模 型

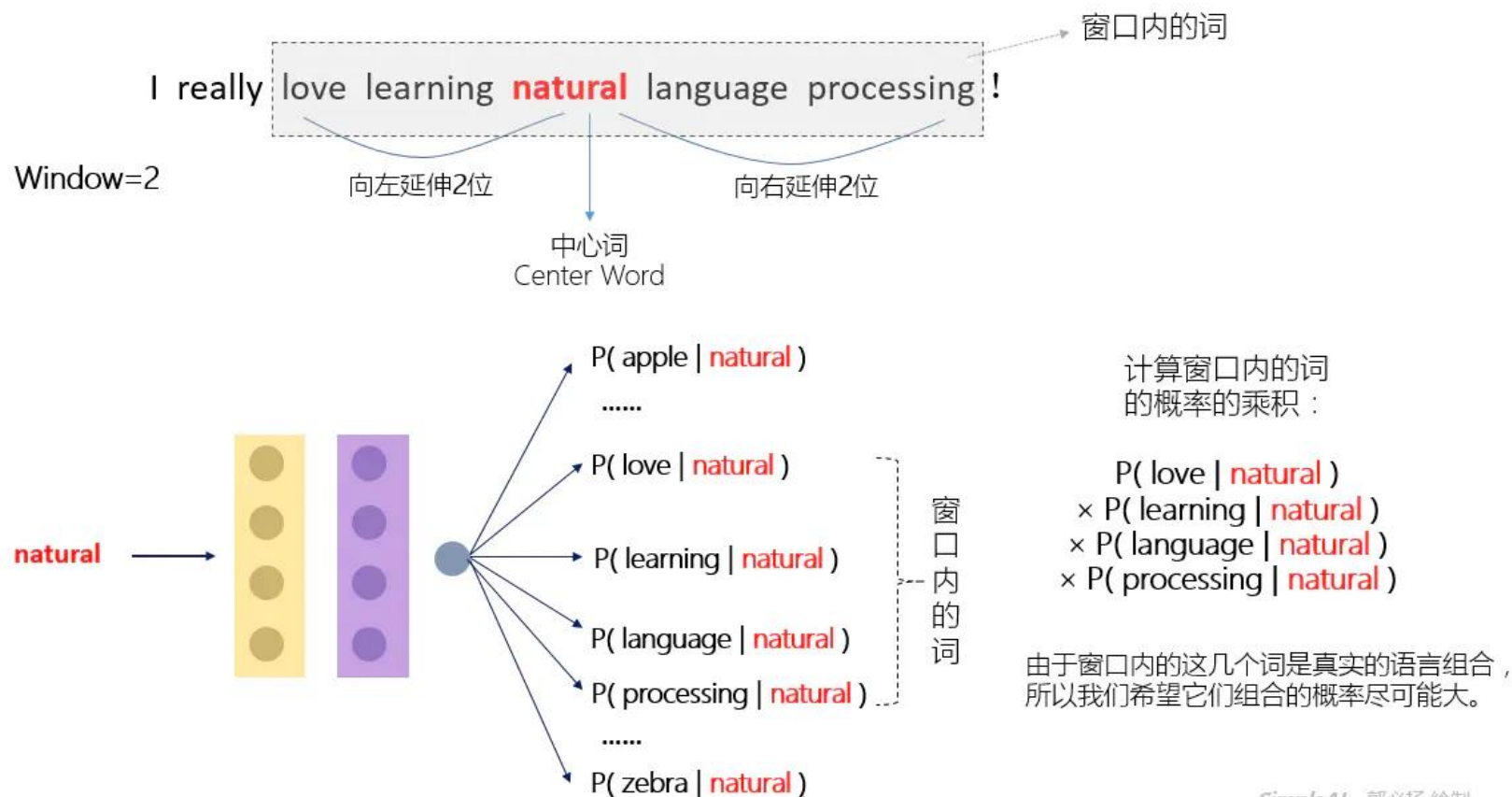
- 基本原则：使语义相近的单词在向量空间中尽可能接近，不相近的单词在向量空间中尽可能拉远。

- 目标函数如下：

$$-d[v(x), v(y)] + d[v(x), v(z)]$$

- 其中， x 和 y 相近； x 和 z 较远
- 通过不断地调整每个单词的词向量，实现该函数的最大化。

词向量模型



训练准则

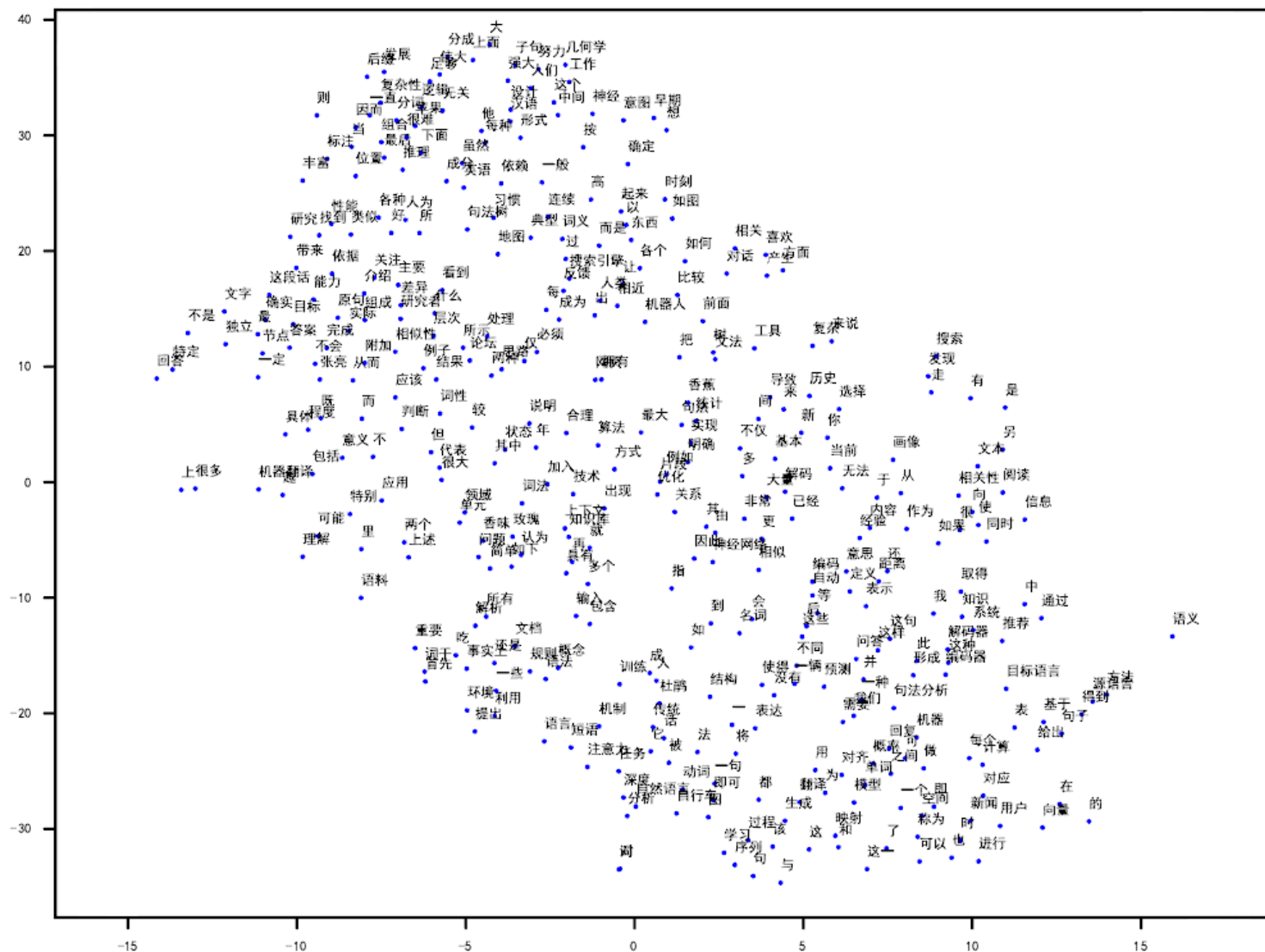
窗口内的词

语言组合

组合概率高

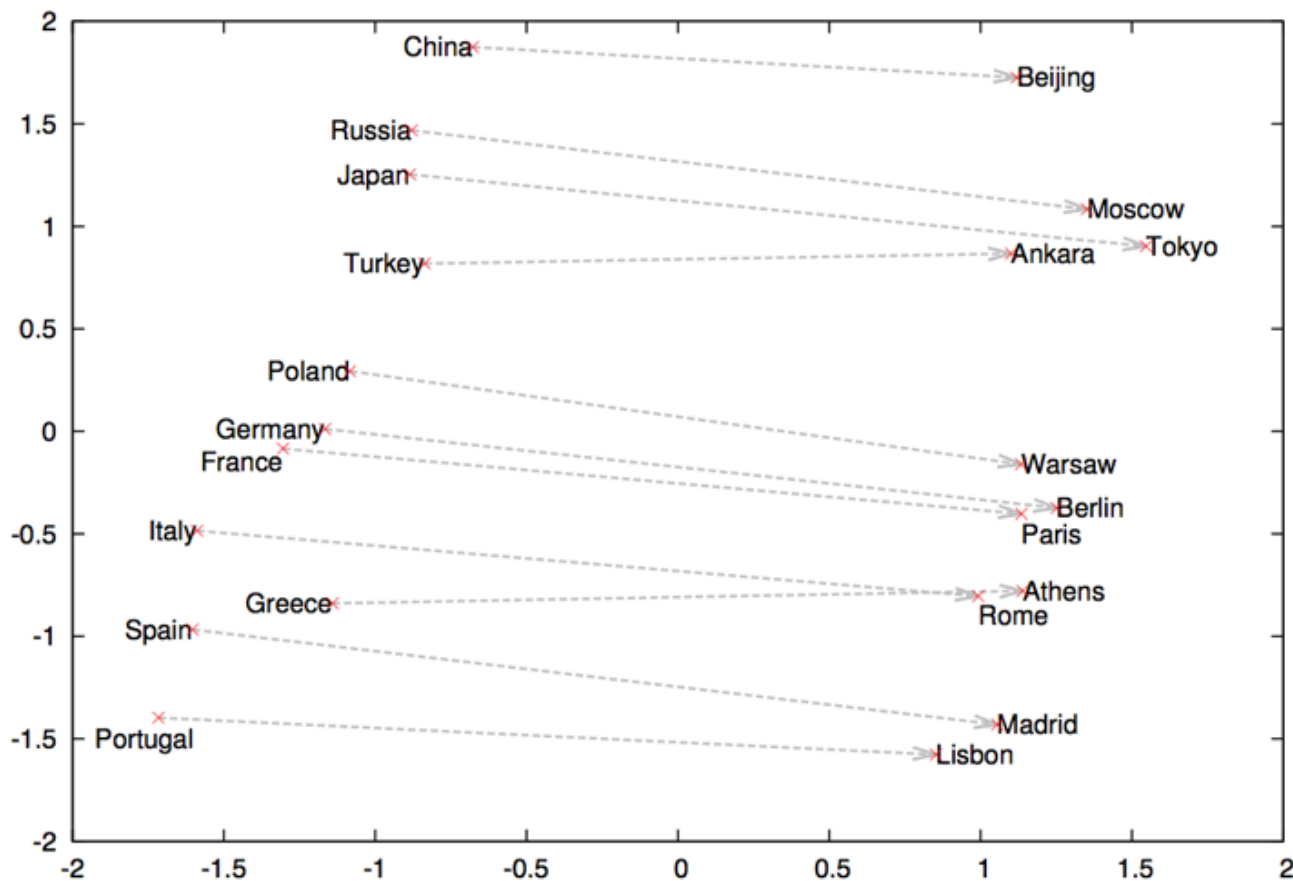
词向量空间

- 具有语义相似性的单词聚在一起了



词向量空间中的隐含关系

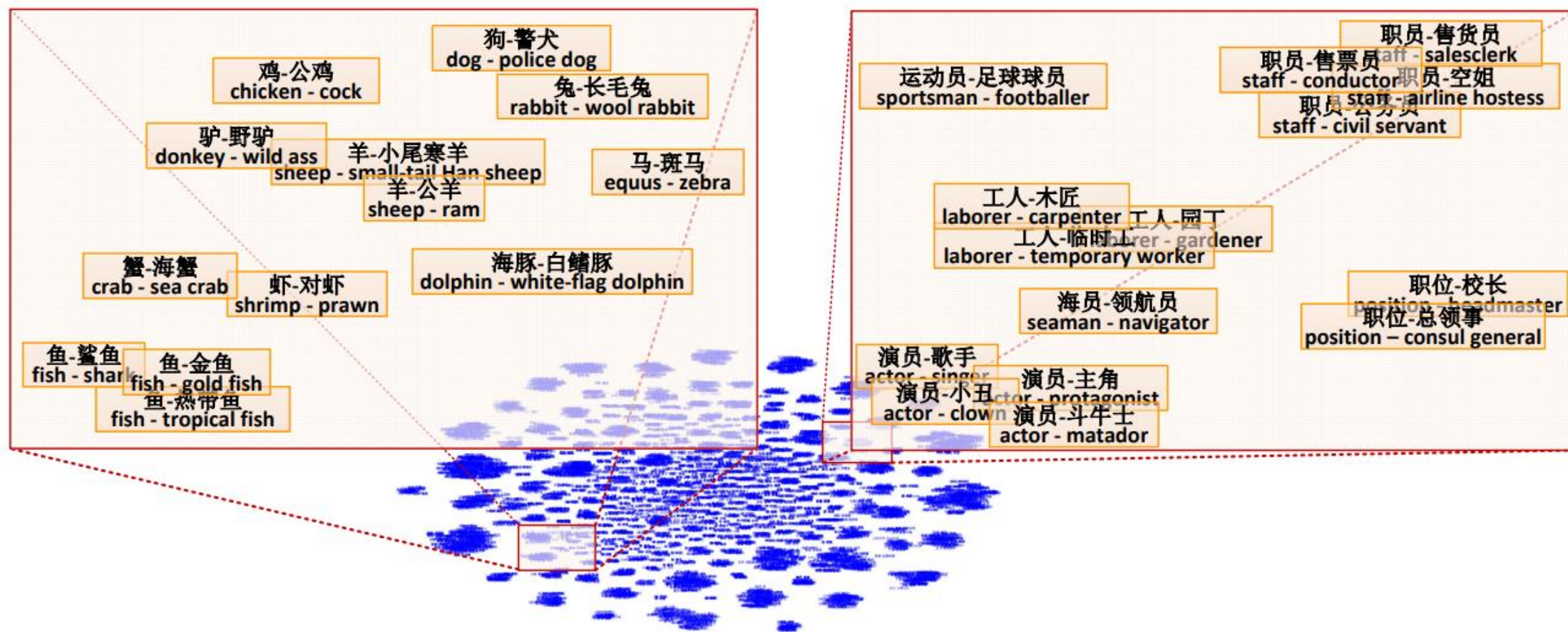
$$W(\text{"China"}) - W(\text{"Beijing"}) \simeq W(\text{"Japan"}) - W(\text{"Tokyo"})$$



连线代表了某种隐含关系（首都）

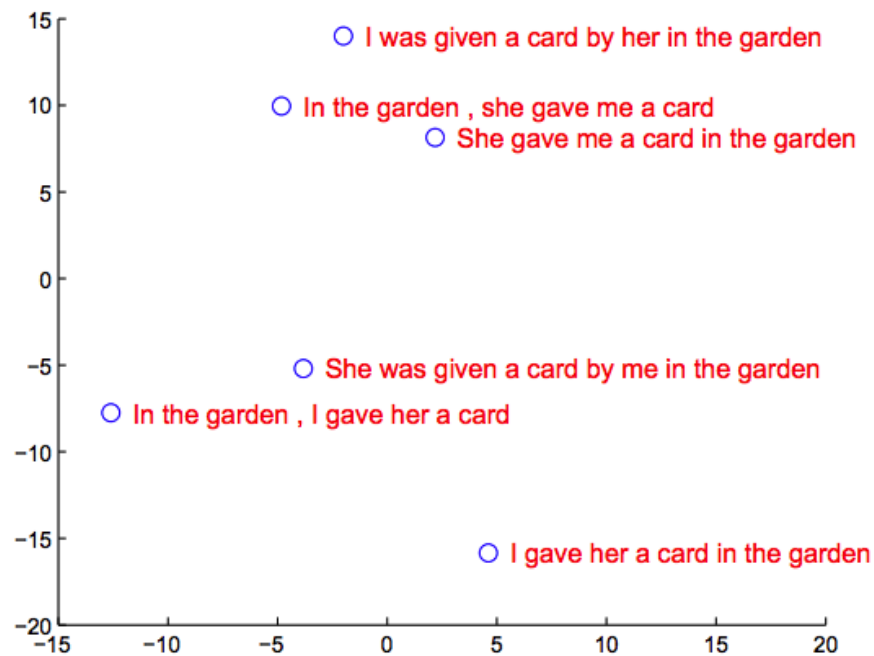
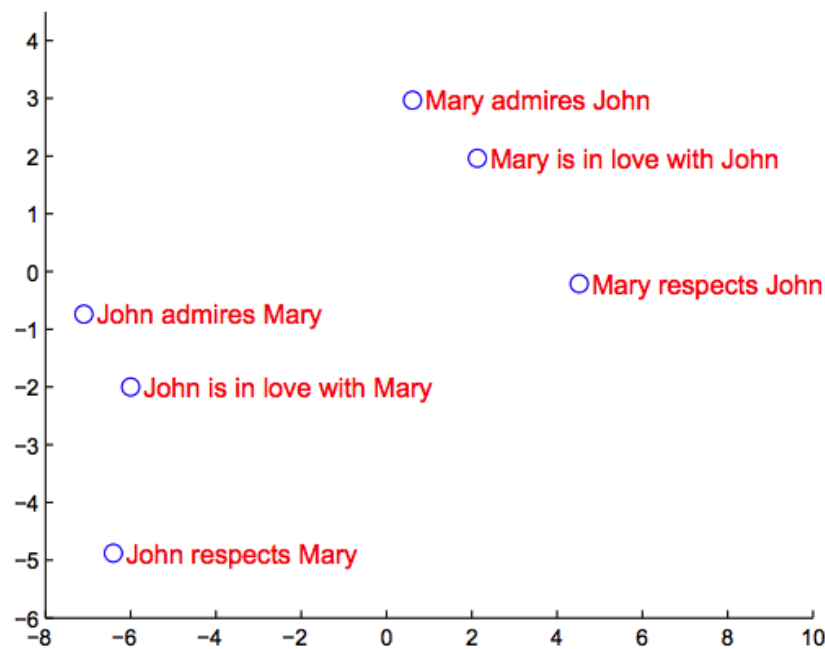
存在相似的隐含关系向量！

词向量空间中的隐含关系



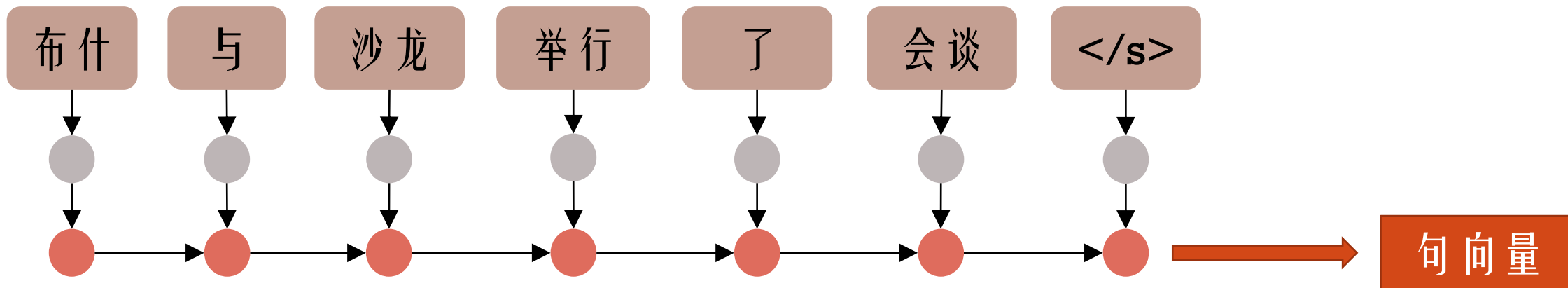
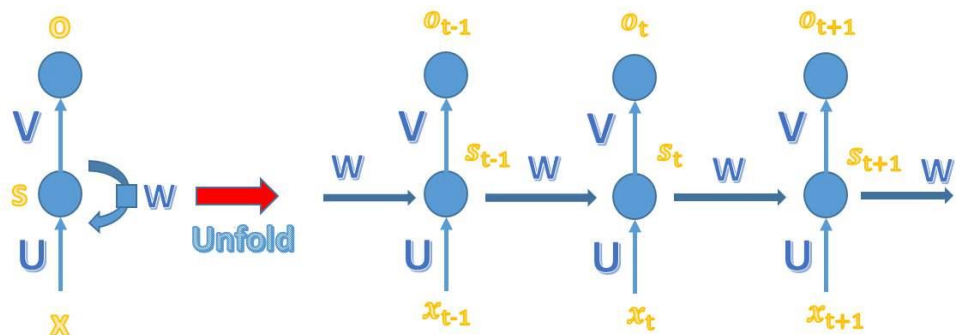
句向量概念

- 由词扩展到句子，对应的词向量扩展到句向量。
- 例如：词向量的平均值来代表句向量。



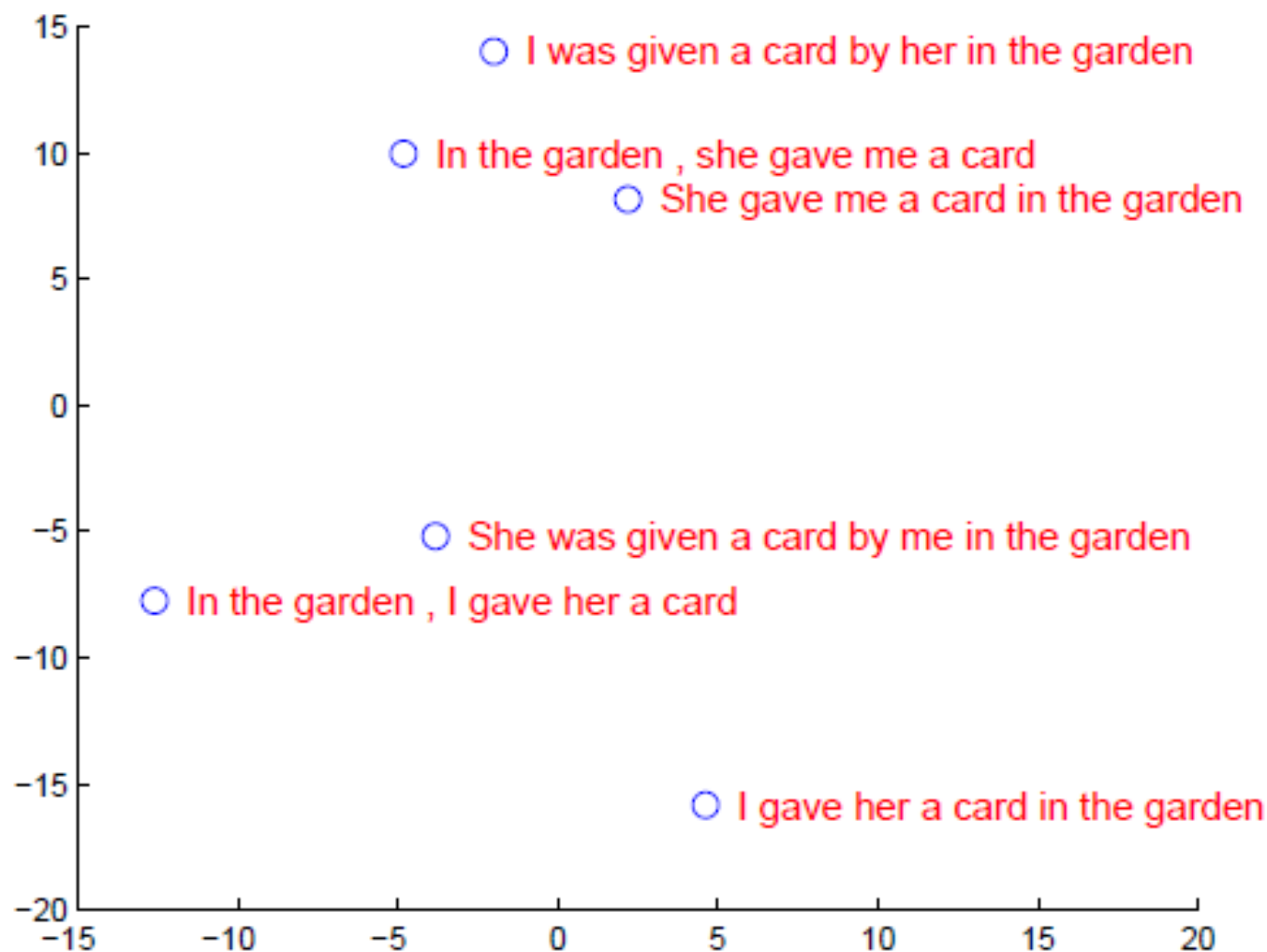
句向量模型

- 循环记忆的神经网络



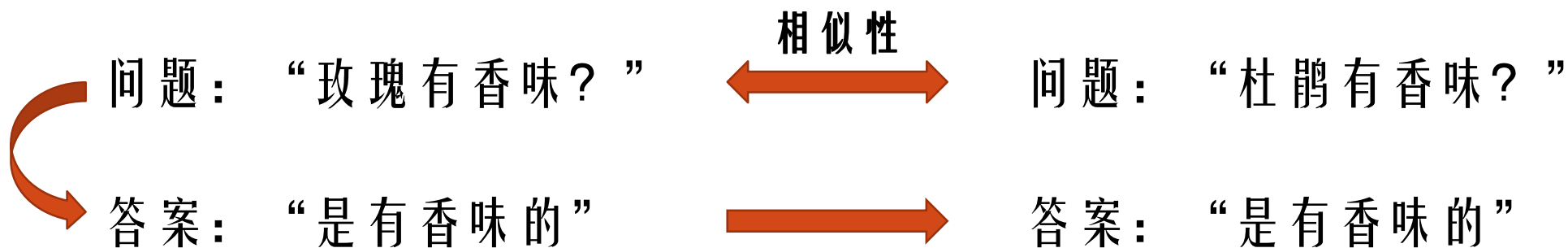
句向量空间

- 具有语义相似性的句子聚在一起



句向量的应用

- 智能问答系统（QA System）
- 数据格式：问题-答案对
- 对于一个问题，在语义空间中寻找与其具有相似性的句向量，从而得到相似问题的答案。



回顾：传统 VS 深度学习

■ 传统方法

■ 优点

- 对句子深入解析，全方位理解
- 基于规则，解析偏差小
- 特定领域、简单句子效果不错

■ 缺点

- 需要丰富的领域知识
- 层次性依赖，易产生错误传导
- 自然语言变化莫测，通用性极差

■ 深度学习方法

■ 优点

- 不需要人为设计知识结构
- 度量简单，易于推理
- 通用性强

■ 缺点

- 语义空间不可解释，黑盒学习
- 难以理解语义来源
- 在特定领域未必有传统方法好

思考

- 词向量和句向量真的可以代表语义吗？
- 它们是如何代表语义的呢？
- 它们又是如何分析语法和句法的呢？

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传 统 语 言 理 解 方 法
- 基 于 深 度 学 习 的 语 言 理 解 方 法
- **机 器 翻 译**
- 语 言 理 解 的 其 他 应 用

通往天堂的高塔

- 《圣经》中记载，人类曾经联合起来兴建能通往天堂的高塔；
- 为了阻止人类的计划，上帝让人类说不同的语言，使人类相互之间不能沟通；
- 计划因此失败，人类自此各散东西。



罗塞塔石碑 (ROSETTA STONE)



古埃及象形文 (Hieroglyphic)
又称为圣书体，代表献给神明



埃及草书 (Demotic)
又称为世俗体，是埃及平民使用文字



古希腊文，代表统治者的语言

自然语言的复杂多样性

I love you

Я люблю тебя

我爱你

당신을 사랑합니다

Eu te amo

Je t'aime

אני אוהב אותך

Ich liebe dich

من شما را دوست دارم

Tôi yêu bạn

Te quiero

Miluji tě

Ti amo

ผมรักคุณ

わたしは、あなたを愛しています

Ik hou van je

Jag älskar dig

如何解决语言不通的障碍？！

机器翻译

- 利用计算机实现自然语言的自动翻译，是自然语言理解中的重要领域

布什

与

沙龙

举行

了

会谈



Bush

held

a

talk

with

Sharon

起源于冷战

- 1947年，瓦伦·韦弗提出了机器翻译的设想。
- 1954年1月7日，IBM在纽约的总部启动了 Georgetown-IBM 实验。IBM 701 计算机有史以来第一次自动将 60 个俄语句子翻译成了英语。

“一位不懂俄语的女孩在IBM的卡片上打出了俄语信息。‘电脑’以每秒2.5行的惊人速度，在自动打印机上迅速完成了英语的翻译。”

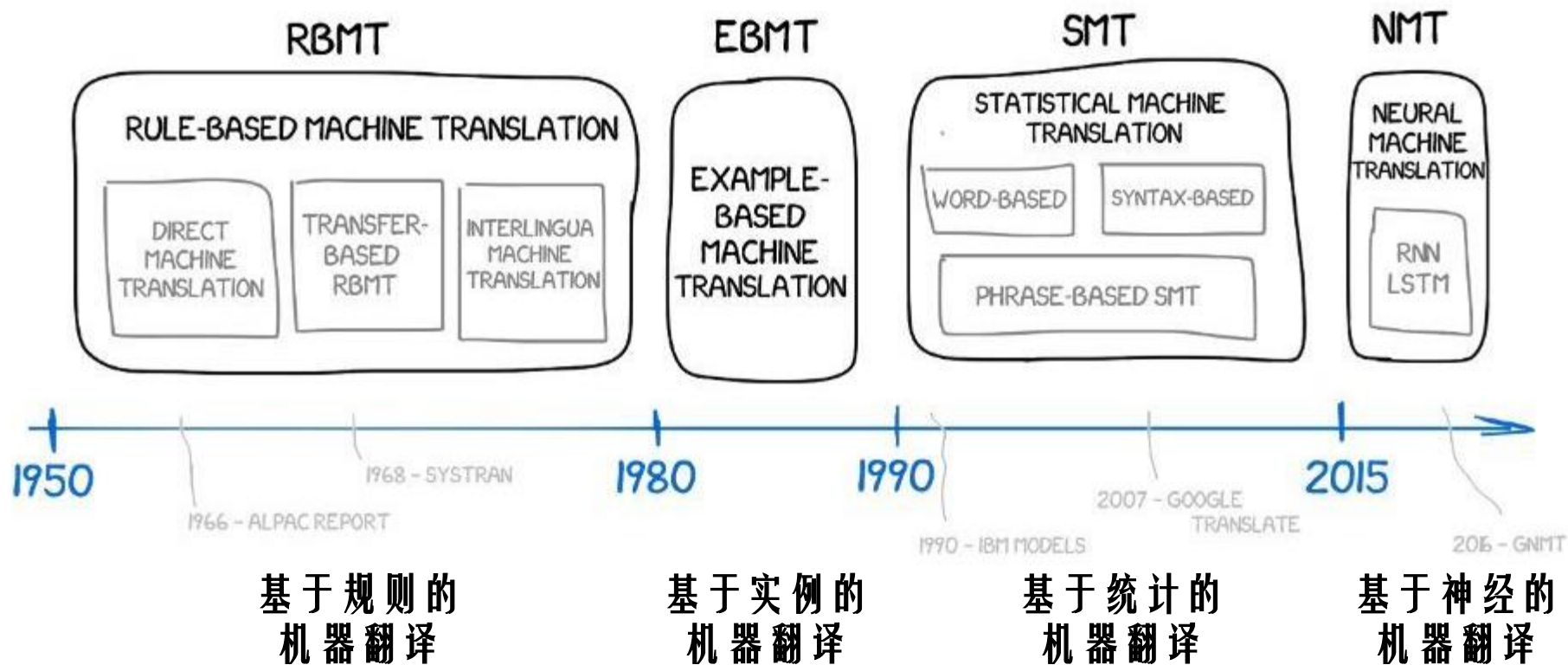
—— IBM 报道



IBM 701

发展历史

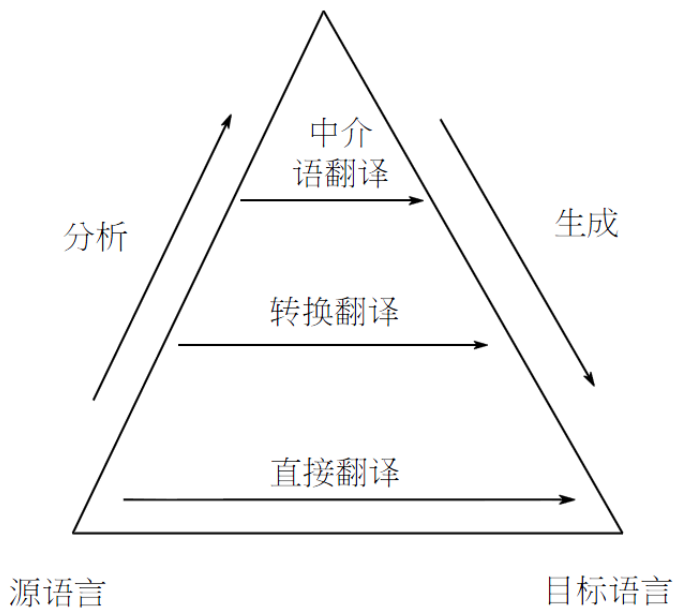
A BRIEF HISTORY OF MACHINE TRANSLATION



基于规则的机器翻译 (RBMT)

- 始于上世纪 70年代
 - 一本双语词典 (俄语->英语)
 - 针对每种语言制定一套语言规则

Vauquois 三角
越底层的翻译单元
越具体；越高层的
翻译单元越抽象



I | WANT | FORTY | KILOGRAMS OF | PERSIMMONS
↓ ↓ ↓ ↓ ↓
ICH | WOLLEN | VIERZIG | KILOGRAMM | PERSIMONEN

直接翻译法

I BOUGHT A SWEET PERSIMMON IN THE STORE
↓ ↓ ↓ ↓ ↓
ICH KAUFTE EINE SÜßE PERSIMONE IM LADEN
↓ ↓ ↓ ↓ ↓
ICH KAUFTE | EINE SÜßE PERSIMONE | IM LADEN

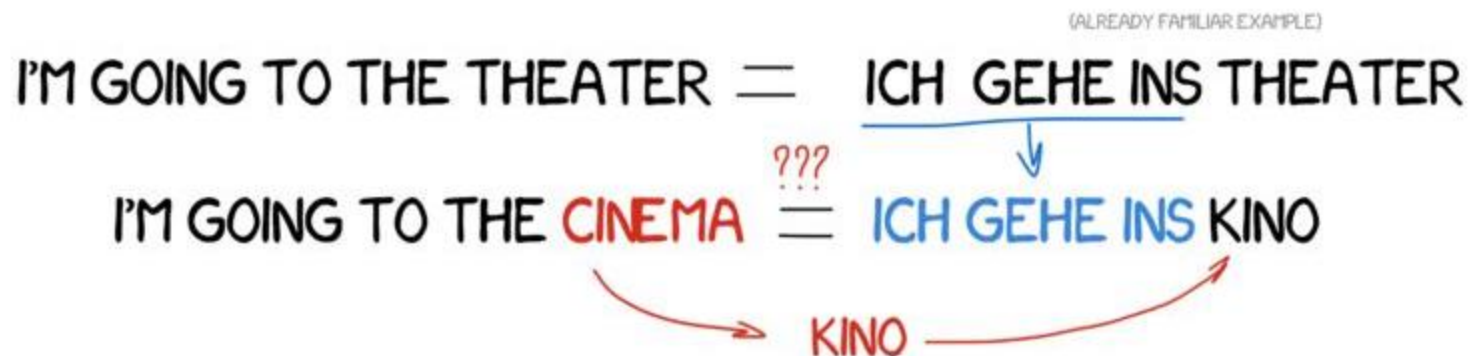
转换翻译法

I | WANT | FORTY | KILOGRAMS OF | PERSIMMONS
↓ ↓ ↓ ↓ ↓
ICH | WILL | VIERZIG | KILOGRAMM | PERSIMONEN

中介翻译法

基于实例的机器翻译 (EBMT)

- 1984年东京大学的长尾真
 - 使用准备好的短语代替重复翻译
 - 保留了句子结构
 - 例子越多，效果越佳
 - 引入了数据驱动的思想！



RMT 存 在 问 题

- 需要大量的语言学知识
- 强加规则，极其蠢笨
- 局限于特定领域（如天气预报）
- 语义是依附上下文环境中的，不同语境的语义不同，翻译理应不同！

I saw a man on a hill with telescope.

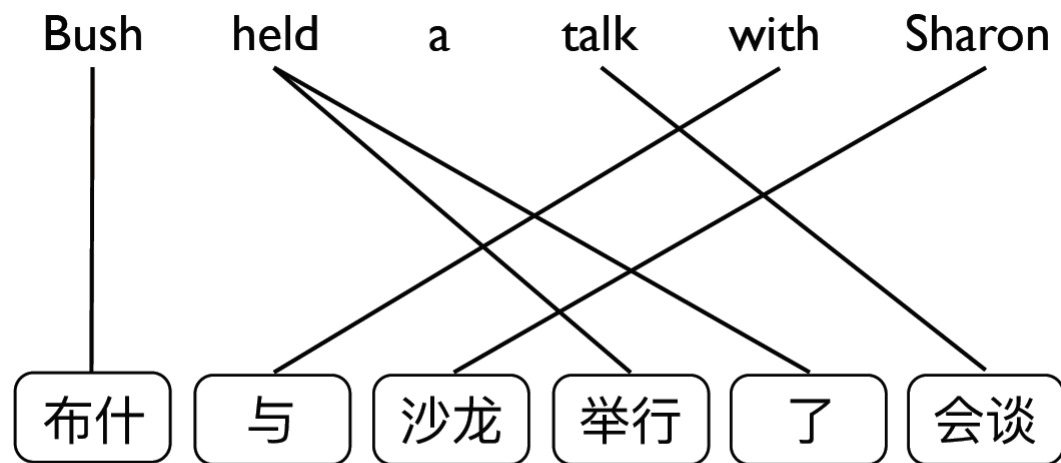
我看到山上有个男人拿着望远镜。

我透过望远镜看到山上站着一个男人。

- 语言是自由的、灵活的、不断发展的，一套规则无法通用！

统计机器翻译（SMT）的诞生

- 1993年 IBM 公司的 Brown 和 Della Pietra
 - 基于词对齐的翻译模型
 - 标志着统计机器翻译（SMT）方法的诞生



(Brown et al., 1993)

SMT的概念

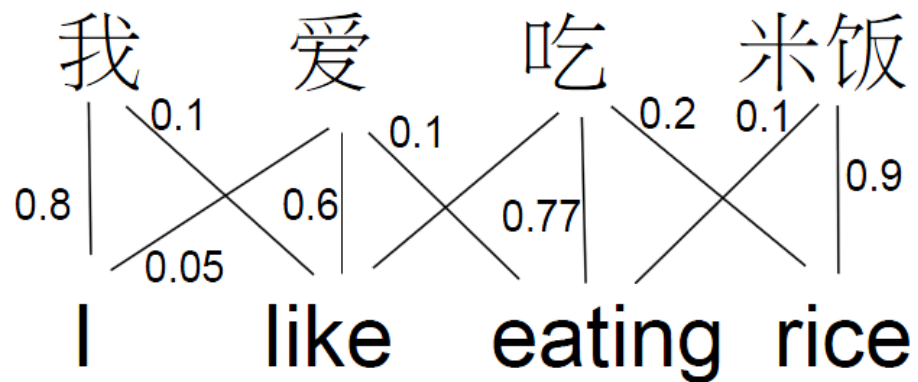
- 基本思路：从双语对应语料中自动学习出两种语言相对应的语言片段，并基于此完成自动翻译。
- 双语对应语料，又称为平行语料，是指两种语言在句子级别的对应语料。

我爱吃鱼 | I love to eat fish

- 极大减轻了机器翻译所需要的资源，推动了技术发展。

SMT 的训练过程

- 发现对应的语言片段，并给每个片段附加一个概率，表示可能性。
- 收集全部可能的对应片段，组成了源语言和目标语言之间的映射表。



我 I 0.1
我 me 0.12
我 am 0.003
走 walk 0.02
走 run 0.003
走 the 0.000001
....

每一行是一个可能的映射。第一列为源语言，第二列为目标语言，第三列为映射的概率。

SMT的翻译过程

- 以“我走去”为例。
- “我走去”基于映射表可选出多个翻译候选，每个候选句对应一个翻译概率。
- 为了判断每个候选句是否合理（是否是“人的话”），需要训练一个语言模型，从而得到每个翻译候选的语言概率。
- 计算翻译概率和语言概率的乘积，最大翻译候选即为最佳翻译。

我	I	0.1
我	me	0.12
我	am	0.003
走	walk	0.02
走	run	0.003
走	the	0.000001
....		

映射表

I	walk	to	0.1
I	walk	there	0.1
I	want	there	0.11
I	run	to	0.03
I	the	me	0.00001
.....			

语言模型

我走去

	翻译概率	语言概率	总概率
I walk there	0.1	0.1	0.01
me walk there	0.05	0.03	0.0015
am run	0.03	0.09	0.0027
I run	0.04	0.11	0.0044
....			

SMT 存 在 问 题

- 明显的“直译”痕迹
- 词语组合生硬，不流畅
- 无法解决长距离调序问题
- 机械式搜索，并不理解句子语义

Chinese

美国总统布什昨天在白宫与以色列总理沙龙就中东局势
举行了一个小时的会谈。✕

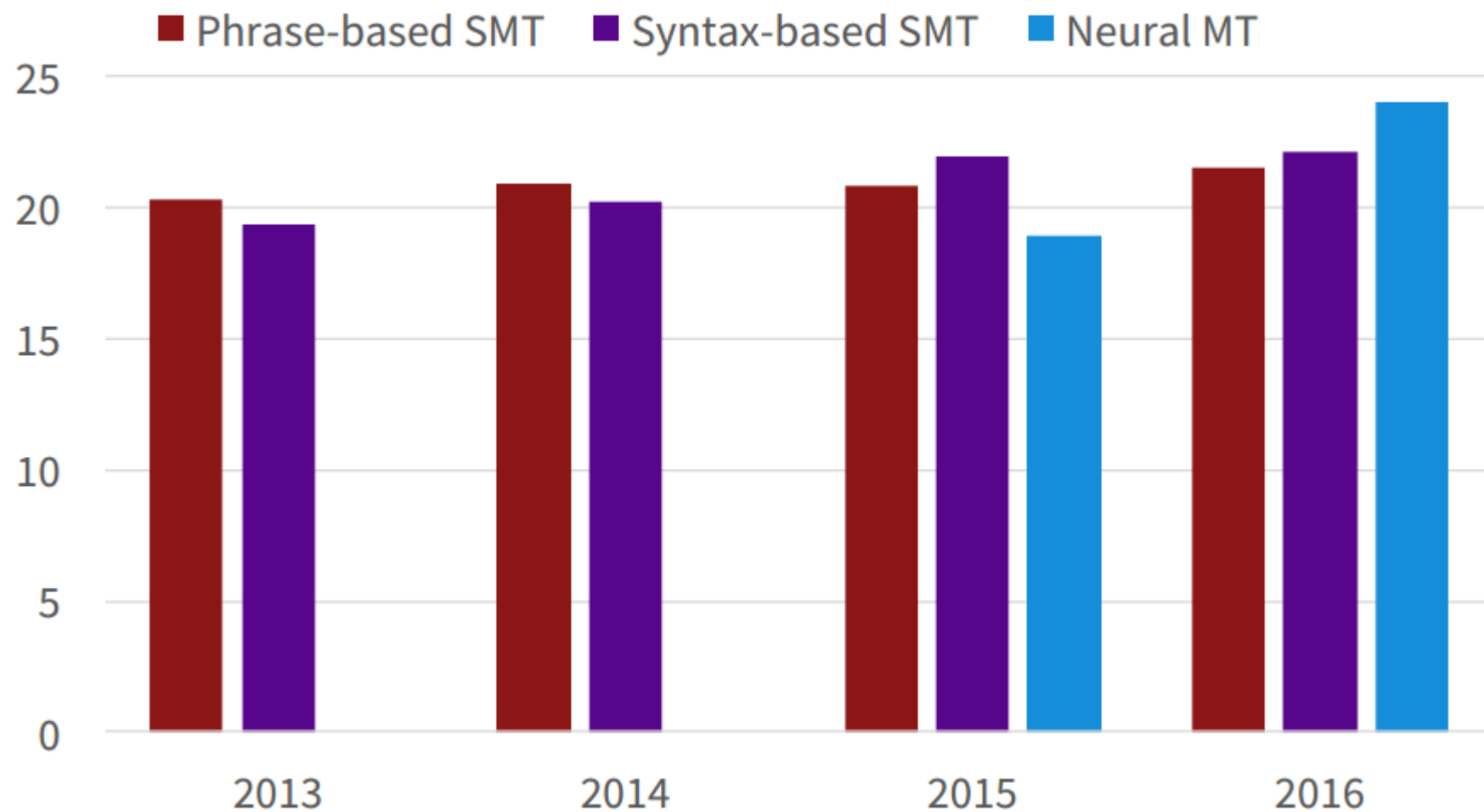
English

Yesterday, U.S. President George W. Bush at the White House with Israeli Prime Minister Ariel Sharon on the situation in the Middle East held a one-hour talks.

深度学习带来新的思路

- 强大的语义理解能力（词向量和句向量）
- 利用深度神经网络实现自然语言之间的映射
- 深度理解，目标驱动

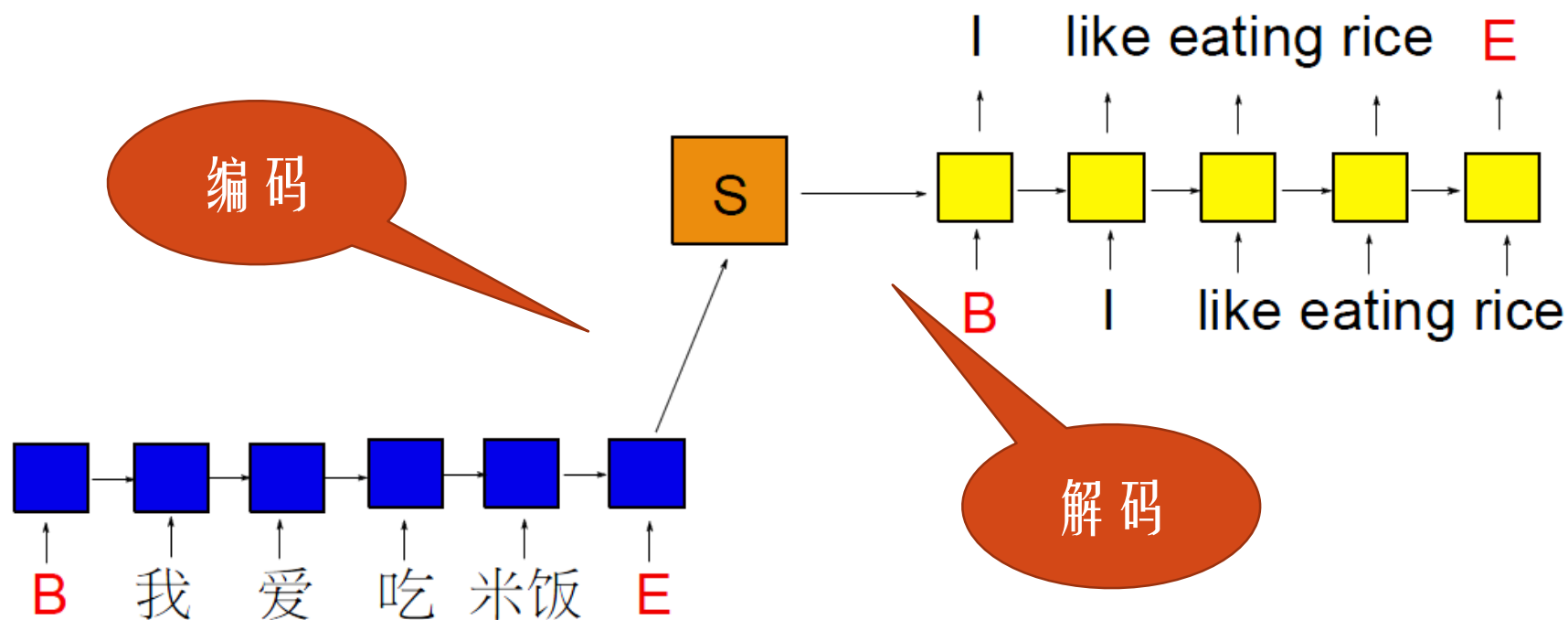
机器翻译方法对比



NMT 逐渐取代 SMT

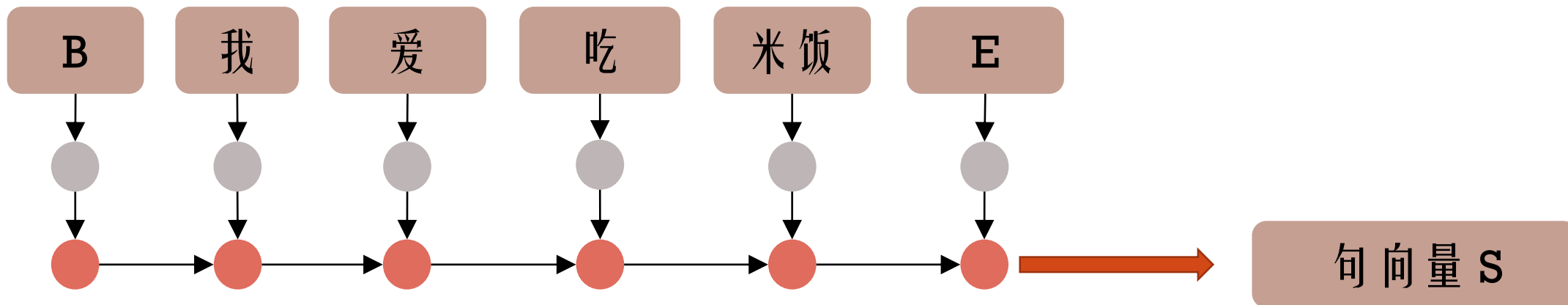
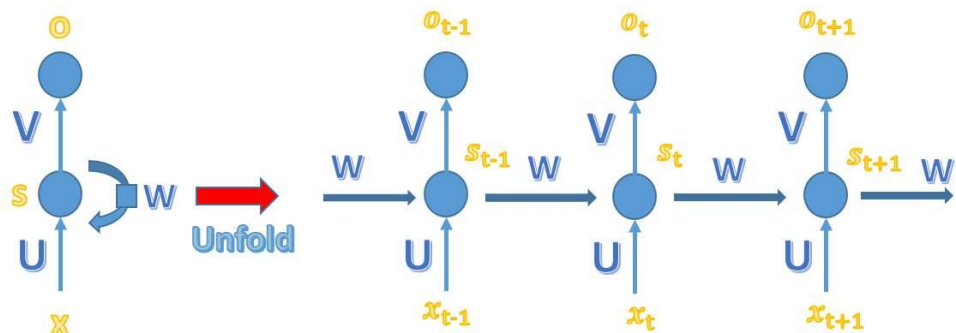
神经机器翻译 (NMT)

- 基础：词向量和句向量的深度语义理解
- 模型：基于编码器-解码器的序列对序列模型



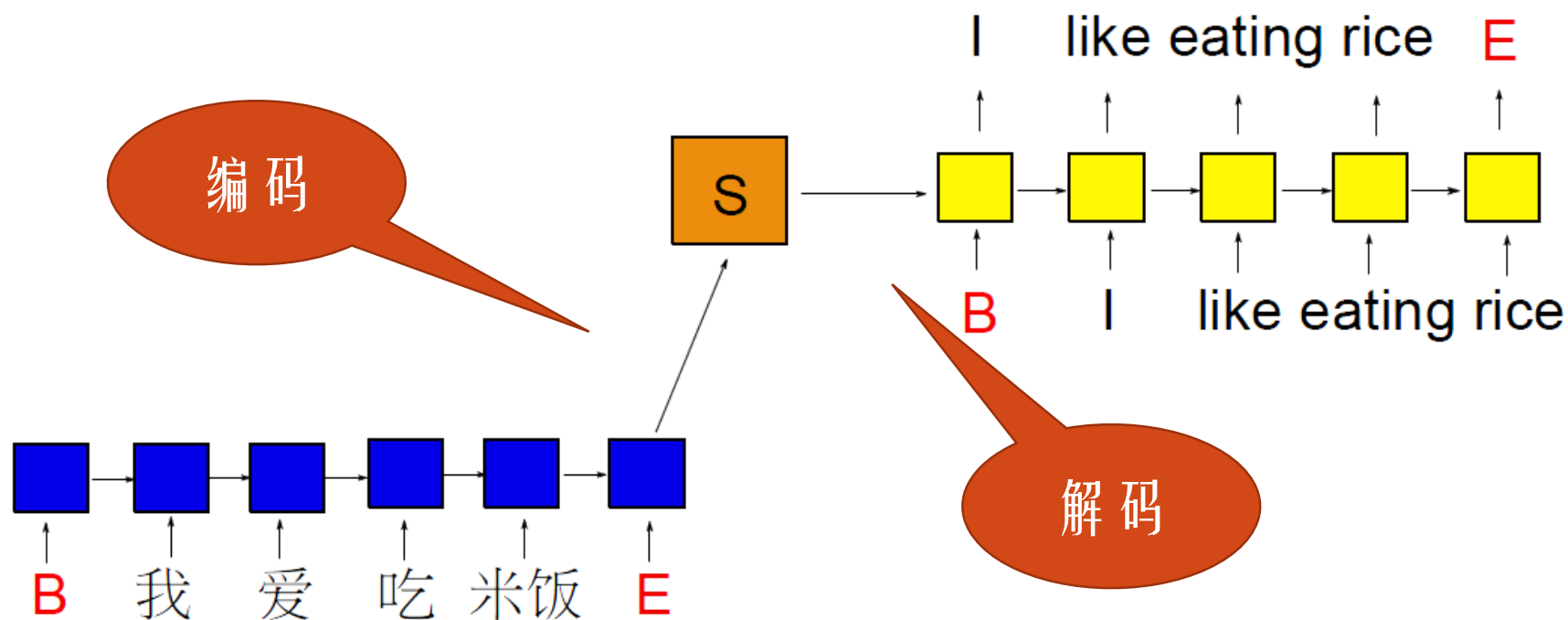
序列模型（同 PPT 49）

- 循环递归神经网络（RNN）



编码-解码

- 编码：待翻译句子被 RNN “压缩” 成一个句子向量 S （语义打包）
- 解码：句子向量 S 通过 RNN “生成” 目标翻译句子（语义重现）



存在问题

- 当句子较长时，句向量 s 的语义表达能力不足



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.

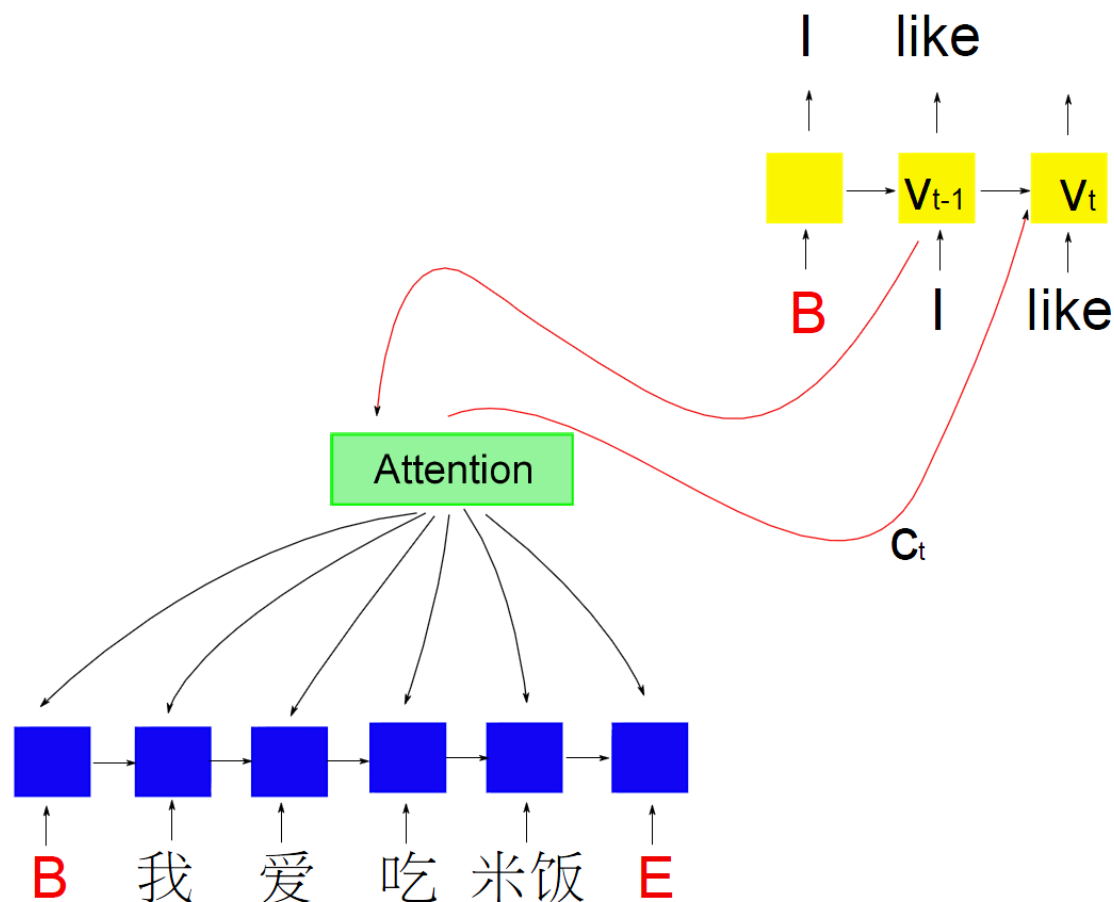


A giraffe standing in a forest with trees in the background.

【注意力机制】
不同位置的关注意度不同

注意力机制

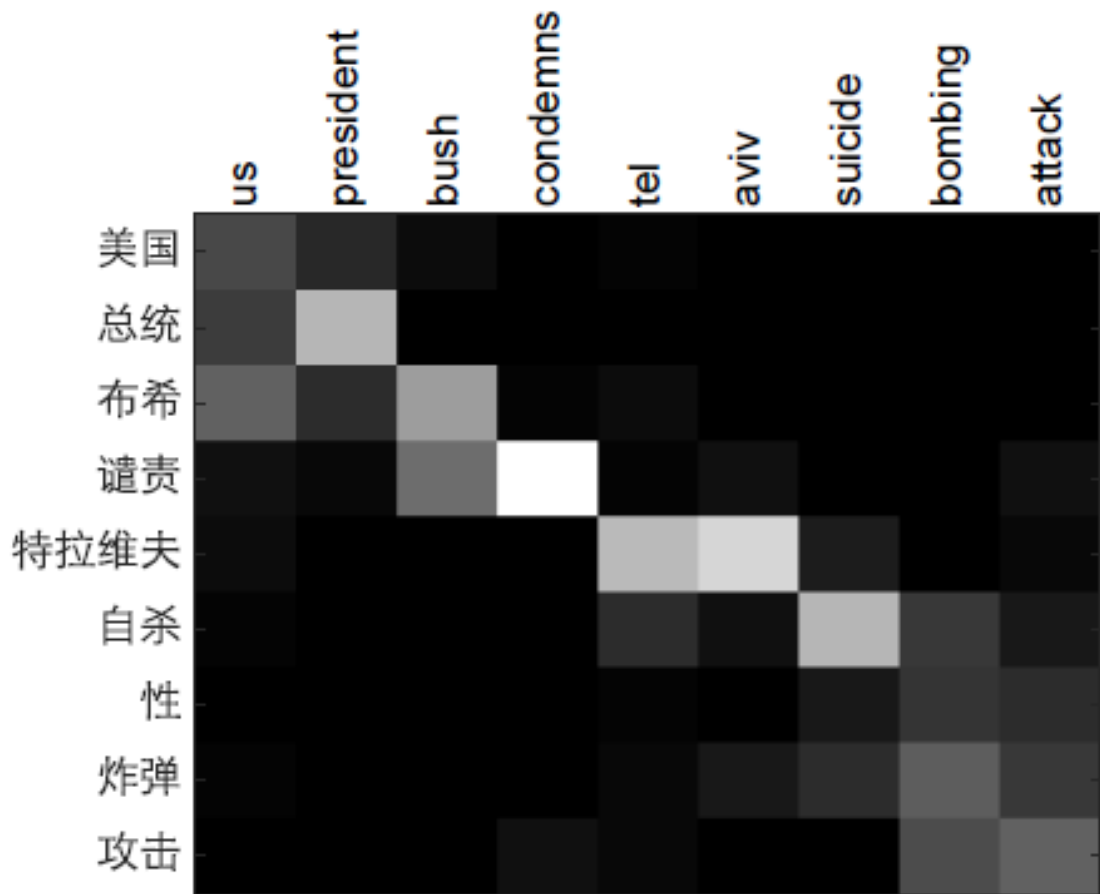
- 存在问题：当句子较长时，句向量 s 的语义表达能力不足
- 解决方案：在原句中寻找当前应该特别关注的位置（例如，特别关注影响当前词的上下文）
- 极大地提升了长句翻译的效果



注意力机制

源语言词与目标语言词的关注度 (集中关注当前词的上下文)

(Bahdanau et al., 2015; Cheng et al., 2016b)



思考

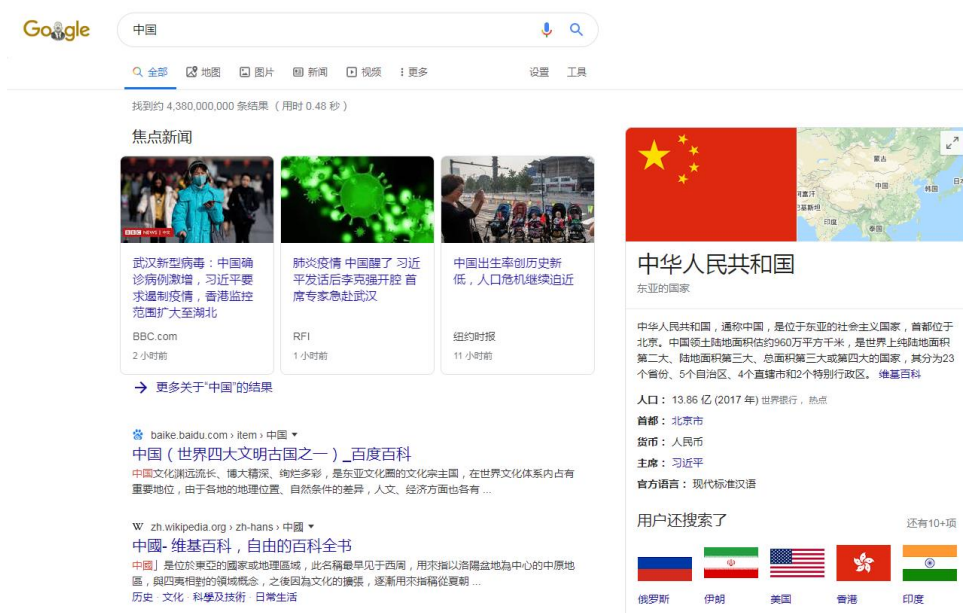
- RBMT、SMT 和 NMT 三者各自的优点和缺点是什么？

目 录

- 引 言
- 人 类 语 言 的 复 杂 性
- 传 统 语 言 理 解 方 法
- 基 于 深 度 学 习 的 语 言 理 解 方 法
- 机 器 翻 译
- 语 言 理 解 的 其 他 应 用

应用一：搜索引擎

- 搜索引擎的发明无疑是近二十年来最伟大的成就之一。
- 在浩如烟海的互联网世界找到自己所需的资源。



Google



Baidu

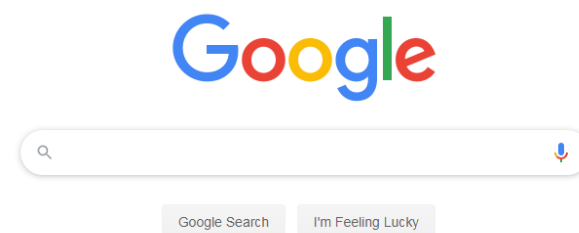
发展历史



最早搜索引擎：1990年的 Archie 系统



1994年 InfoSeek 浏览器



1998年 Google 正式成立



第一个面向网页的搜索引擎：1993年的 World Wide Web Wanderer



1994年 YaHoo 搜索引擎



2000年百度正式成立

搜索技术的发展

- 向量空间法
 - 每个网页表示为一个文档向量
 - 该向量代表了网页中每个单词的重要程度（词频）
 - 比较用户输入的文档向量和数据库中每个网页的文档向量

搜索技术的发展

■ PageRank

- Google 的创始人拉里·佩奇和谢尔盖·布林于1998年发明的
- 充分利用网页之间的相关性
- 这种相关性包括网页超链接、浏览网页时不同网页之间的跳转关系等
- 利用这种相关性计算网页间的相似度

How PageRank Works (A Simplified View)

PageRank is divided equally between the total number of links on a page.



Source: <https://ahrefs.com/blog/google-pagerank/>

ahrefs

搜索技术的发展

- 用户意图搜索

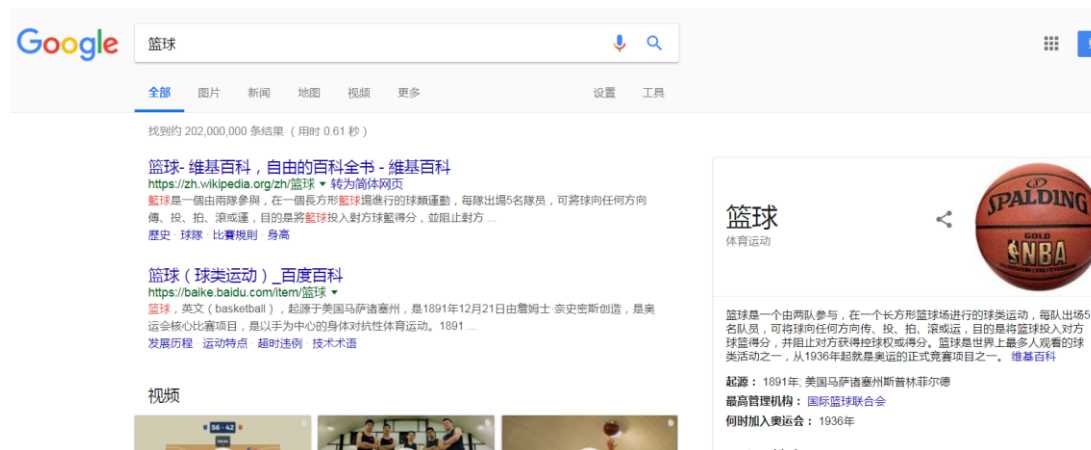
- 网页列表

- 图片、视频等丰富的信息

- 人性化搜索

- 分析用户搜索历史

- 调整搜索结果



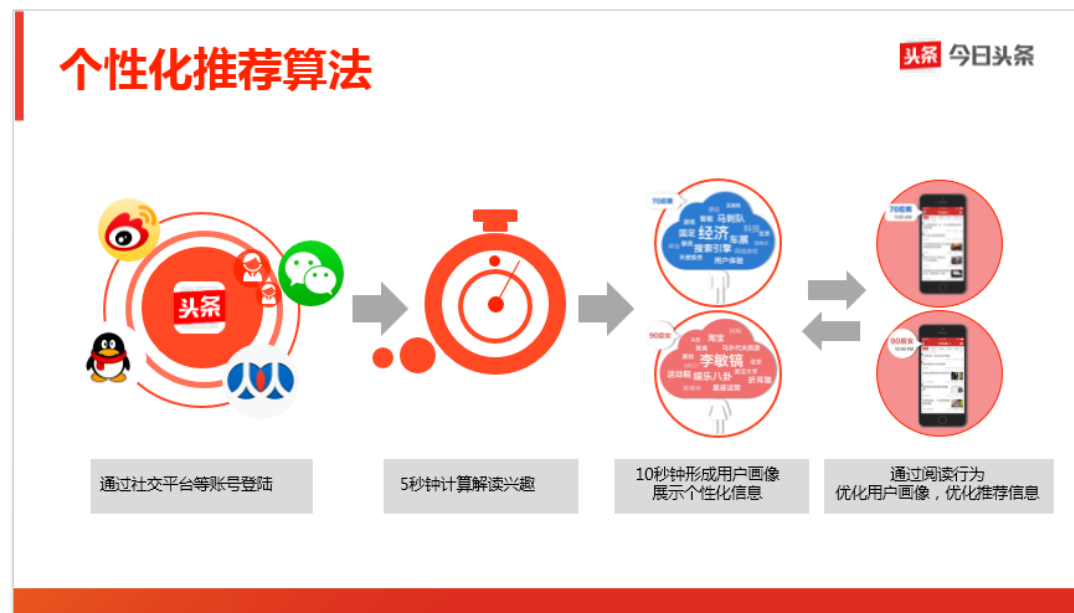
应用二：推荐系统

- 推荐是指我们在搜索某一内容时，系统自动为我们推荐相关的其他内容。

发现好货 > 发现品质生活



京东购物推荐



今日头条 APP 新闻推荐

以新闻推荐为例

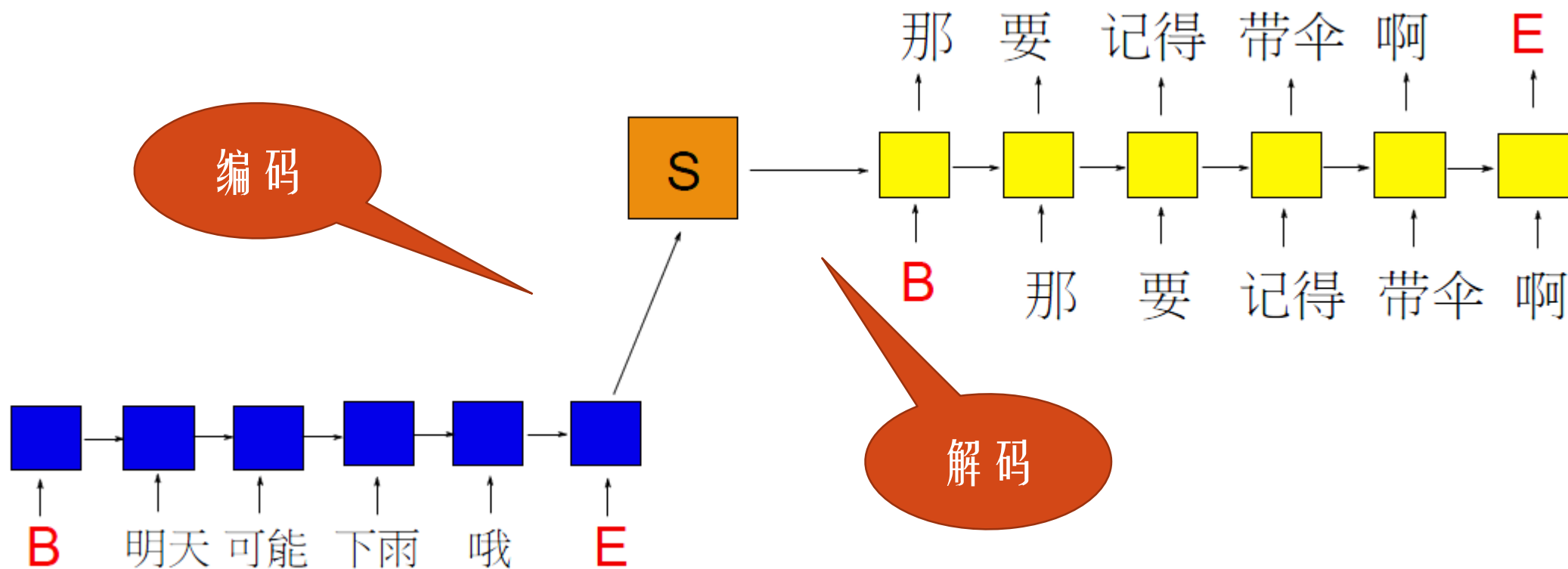
- 新闻向量化 和用户向量化（用户画像）
- 用户与新闻的相关性
 - 内容推荐
- 新闻与新闻的相关性
 - 扩展推荐
- 用户与用户的相关性
 - 交叉推荐

应用三：会话机器人

- 会话机器人：可以和人自由聊天的机器人，如 Siri、微软小冰、小爱同学



基于序列对序列模型的会话系统



智能问答系统

Post	意大利禁区里老是八个人... 太夸张了 There are always 8 players at the Italian restricted area. Unbelievable!
Response1	我是意大利队的球迷，等待比赛开始。 I am a big fan of the Italy team, waiting for the football game to start
Response2	意大利的食物太美味了 Italian food is absolutely delicious.
Response3	太夸张了吧! Unbelievable!
Response4	哈哈仍然是0:0 。还没看到进球。 Ha!Ha!Ha! it is still 0:0, no goal so far.
Response5	这正是意大利式防守足球。 This is exactly the Italian defending style football game.

AI VS AI: 机器人对话



思考

- 展望未来的语言理解技术发展和应用场景

The end