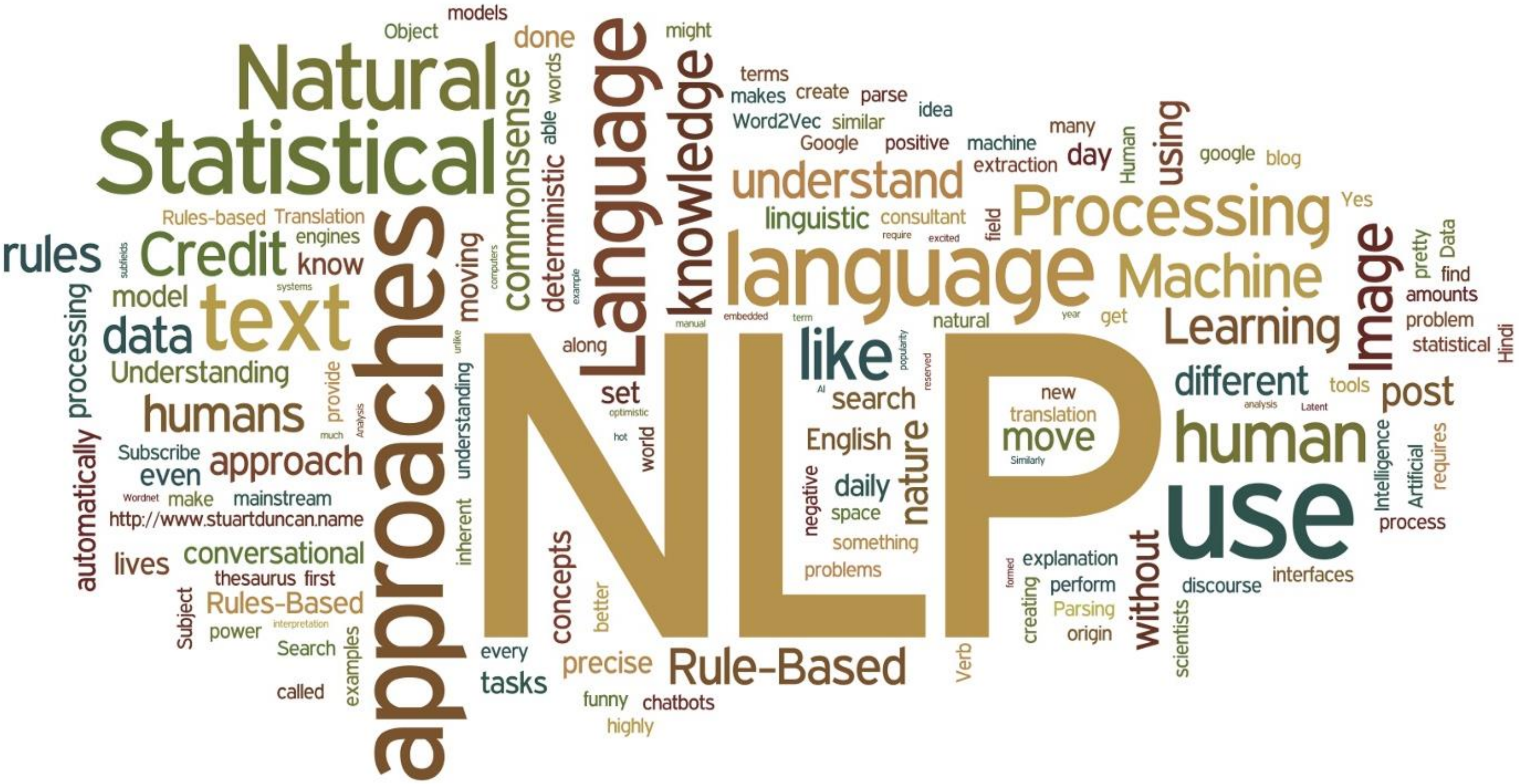
自然语言处理

赵云蒙
华东理工大学 信息科学与工程学院
能源化工过程智能制造教育部重点实验室
2022-2023 第一学期

第一章 绪论

1.1 问题的提出

1.1 问题的提出

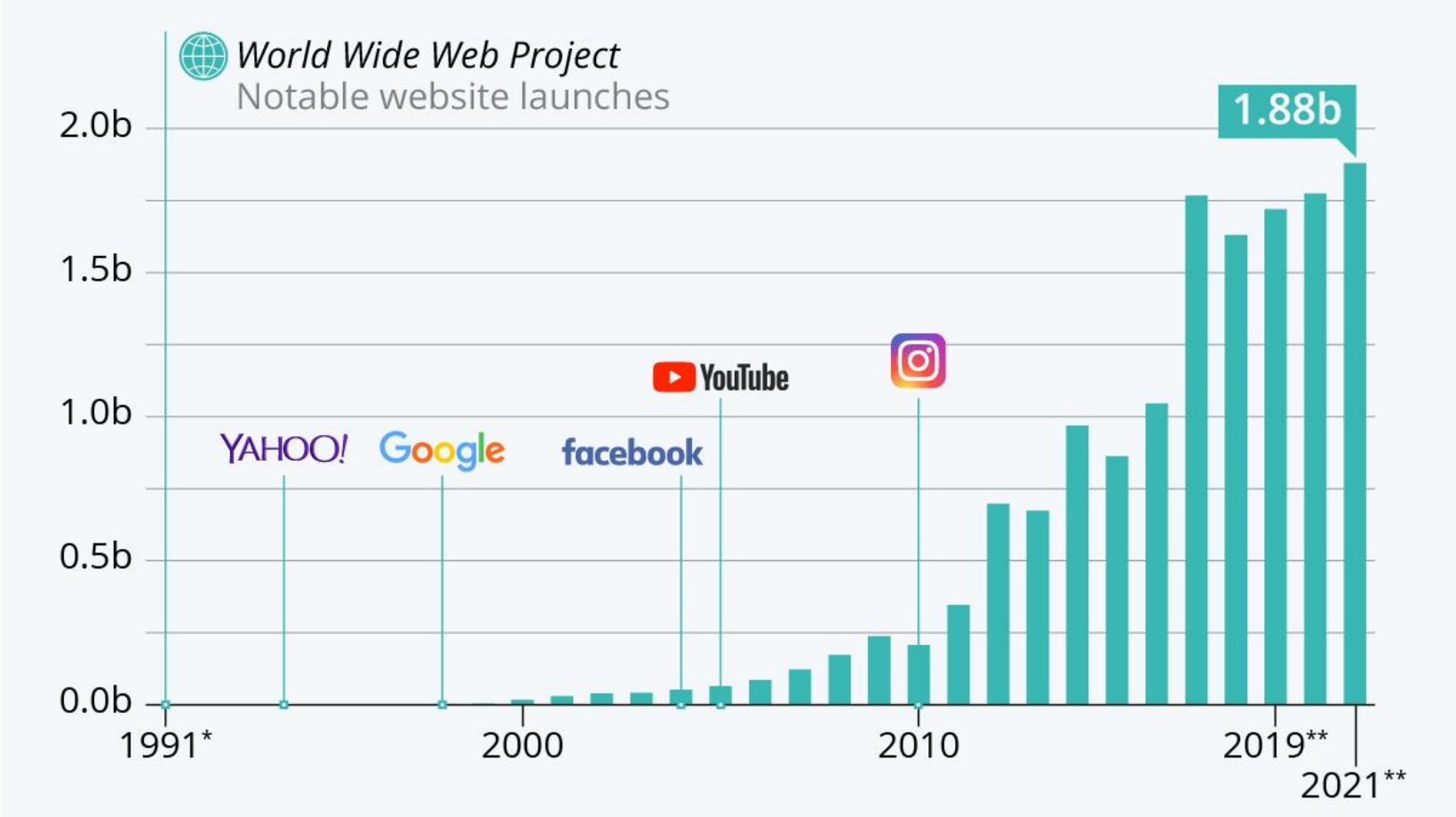

NLP
language
Language
approaches
Statistical
Natural
Processing
Machine
Learning
use
human
text
Credit
knowledge
understand
like
Rule-Based
humans
approach
data
rules
Image
different
post
move
search
English
nature
daily
commonsense
deterministic
concepts
without
model
Understanding
conversational
Rules-Based
Rules-based
Translation
engines
know
models
Object
done
might
terms
makes
create
parse
idea
Word2Vec
similar
Google
positive
machine
many
extraction
day
Human
using
google
blog
Yes
linguistic
consultant
field
pretty
Data
find
amounts
problem
statistical
Hindi
natural
get
tools
Intelligence
Artificial
requires
process
interfaces
discourse
explanation
perform
Parsing
origin
creating
scientists
Verb
problems
something
space
negative
new
translation
set
along
world
inherent
understanding
moving
words
able
better
precise
tasks
every
funny
chatbots
highly
processing
automatically
Subscribe
even
make
mainstream
http://www.stuartduncan.name
lives
thesaurus
first
Subject
power
Search
called
examples
provide

1.1 问题的提出

- 任意时间、任意地点、任意语言的自由通讯无时无刻不在改变着人们的思维方式和生活方式
- 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
- 2008年1月中国互联网络信息中心（CNNIC）发布的《第21次中国互联网络发展状况统计报告》表明，中国互联网上有87.8%的网页内容是文本表示的
- 面对文本大数据，我们面临怎样的机遇和挑战?

1.1 问题的提出

- 网络信息检索市场前景广阔
  - 全世界网站数量正以指数速率增长



  - 中文网页检索的最高准确率不足**40%**

1.1 问题的提出

- 全世界正在使用的语言有**1900**多种
- 随着社会全球化时代的到来，机器翻译市场潜力巨大：
  - 文化
  - 商贸
  - 旅游
  - 学术
  - 体育
  - ……

- 跨语言通讯和信息获取技术具有重要的用途

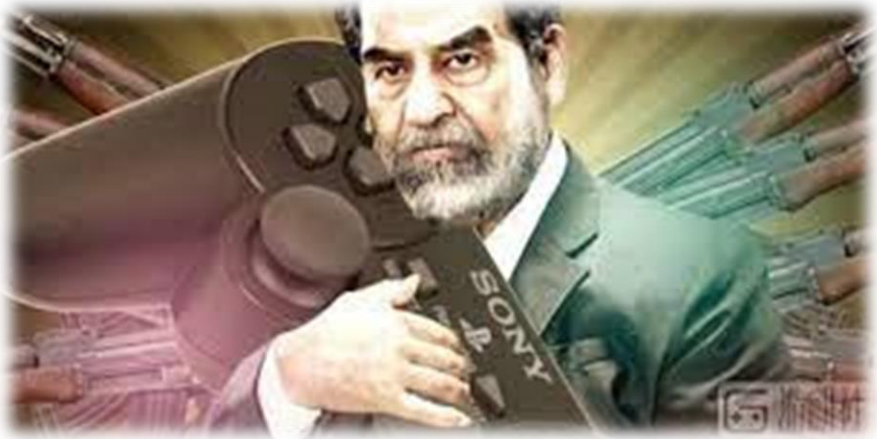
1.1 问题的提出

Threema

Telegram

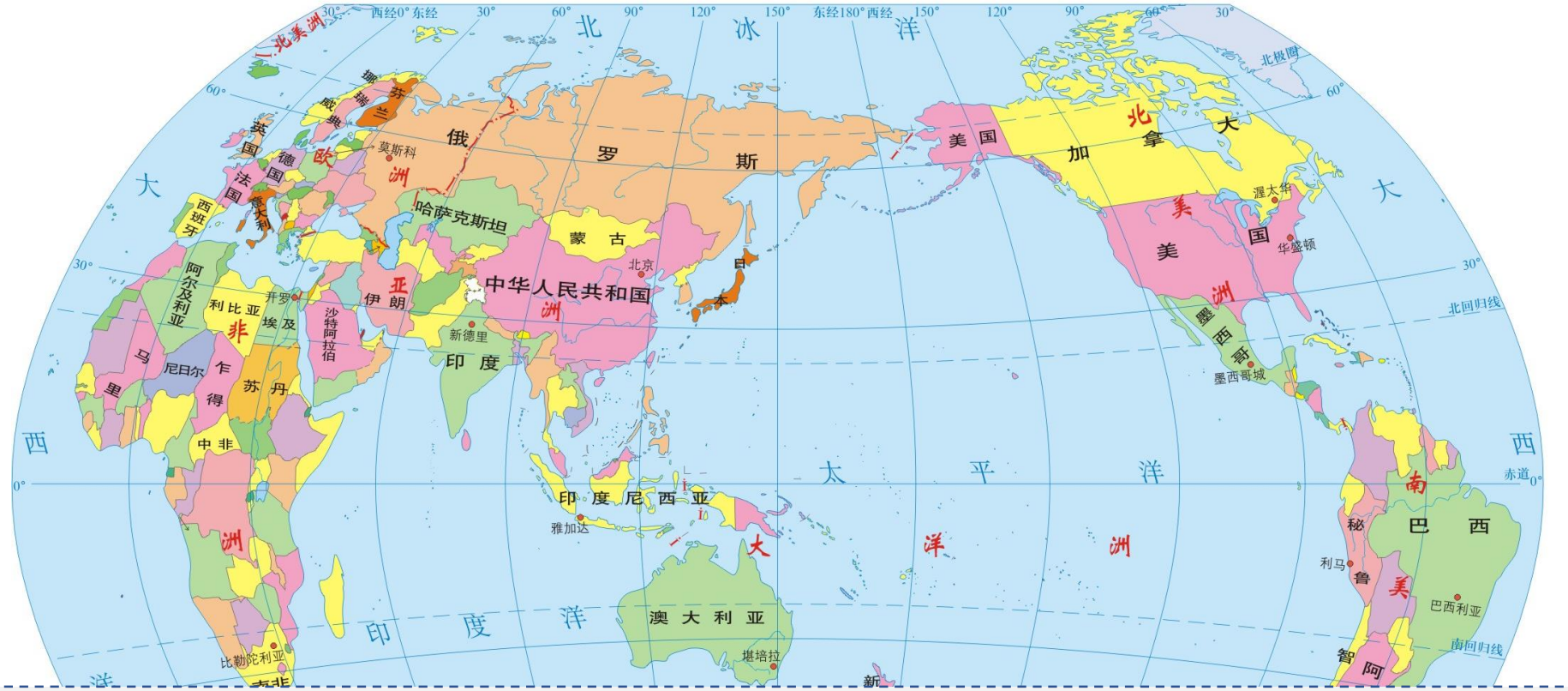
Signal

WhatsApp



利用网络组织犯罪，已成为犯罪活动的新特点

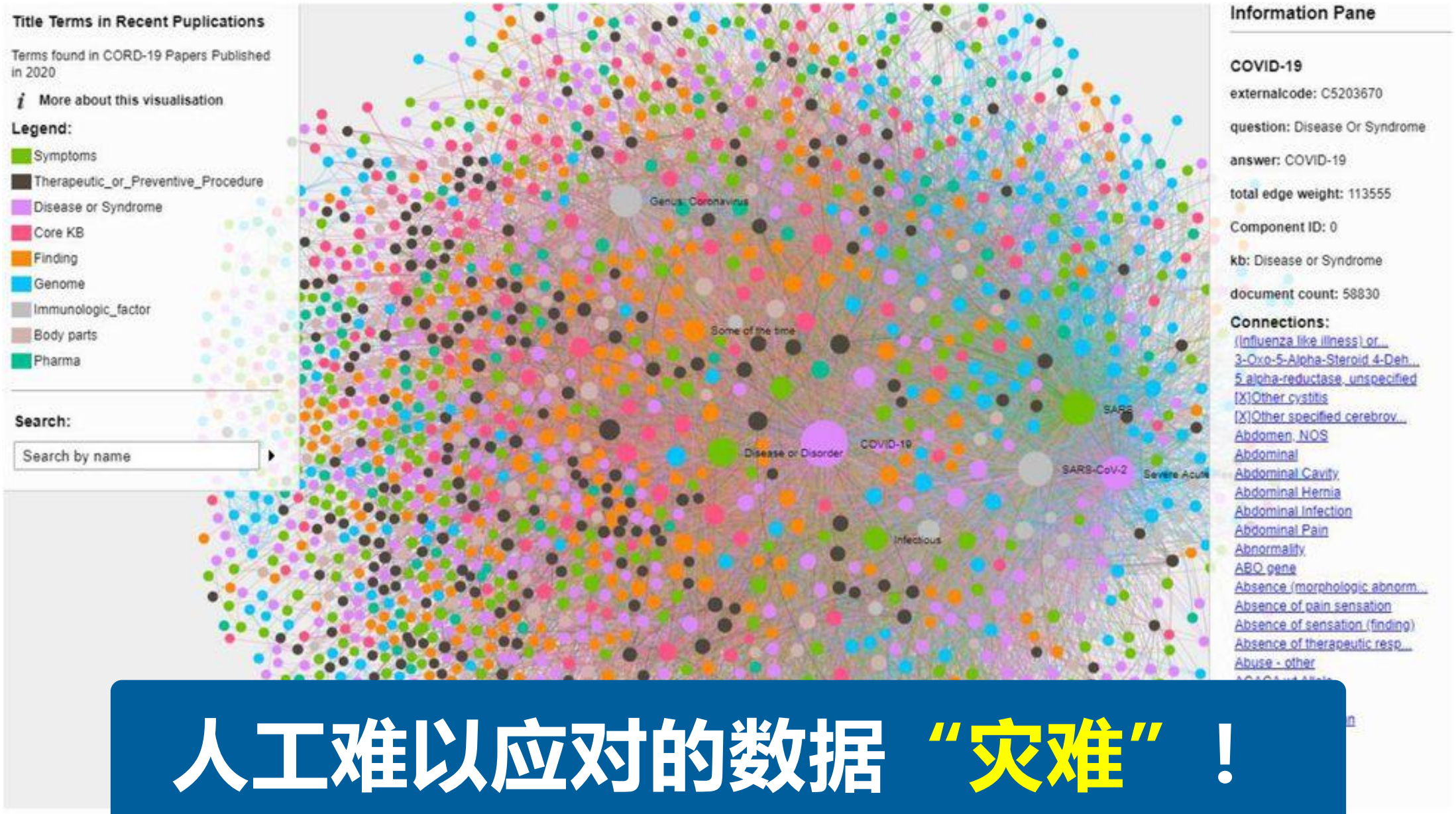
1.1 问题的提出



人们处在不同的国家，使用不同的语言，在不同的地方发表过不同的言论(专著、论文、博客、网页等)，千丝万缕的关系将他们联系在一起，构成一个特定的社会网络。

如何发现或挖掘这种网络？如何确定不同的实体、事件和知识之间的关联？

1.1 问题的提出



人工难以应对的数据“灾难”！

1.1 问题的提出

- 如何让计算机能够自动或半自动地理解自然语言文本，懂得人的意图和心声？
- 如何让计算机实现海量语言文本的自动处理、挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务？

自然语言处理

Natural Language Processing, NLP

1.2 基本概念

1.2 基本概念

- 语言学 vs. 语音学
- 自然语言理解 vs. 自然语言处理

  vs. 计算语言学

  vs. 中文信息处理

1.2 基本概念

✦定义1-1：语言学 Linguistics

语言学是指对语言的科学研究。

—戴维•克里斯特尔，《现代语言学词典》，1997

研究语言的本质、结构和发展规律的科学。

—商务印书馆，《现代汉语词典》，2016

语音和文字是语言的两个基本属性。

1.2 基本概念

✦定义1-1：语言学 Linguistics

作为一门纯理论的学科，语言学在近期获得了快速发展，尤其从上个世纪60年代起，已经成为一门知晓度很高的广泛教授的学科。包括：历时语言学 (diachronic linguistics) 或称历史语言学 (historical linguistics)、共时语言学 (synchronic linguistics)、描述语言学 (descriptive linguistics)、对比语言学 (contrastive linguistics)、结构语言学 (structural linguistics)等等。

1.2 基本概念

✦定义1-2：语音学 Phonetics

研究人类发音特点，特别是语音发音特点，并提出各种语音描述、分类和转写方法的科学。

包括: (1)发音语音学(articulatory phonetics)，研究发音器官如何产生语音；(2)声学语音学(acoustic phonetics)，研究口耳之间传递语音的物理属性；(3)听觉语音学(auditory phonetics), 研究人通过耳、听觉神经和大脑对语音的知觉反应。

—戴维•克里斯特尔，《现代语言学词典》，1997

1.2 基本概念

✦ **问题：**

语音学究竟是一门独立的学科还是应视为语言学的一个分支呢?

复数的语言科学（Linguistic sciences）

1.2 基本概念

✦定义1-3：计算语言学 Computational Linguistics

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）

✦定义1-4：自然语言理解

Natural Language Understanding, NLU

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）

1.2 基本概念

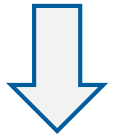
- **关于“理解”的标准**

- **如何判断计算机系统的智能?**

计算机系统的表现（act）如何?

反应（react）如何?

相互作用（interact）如何?

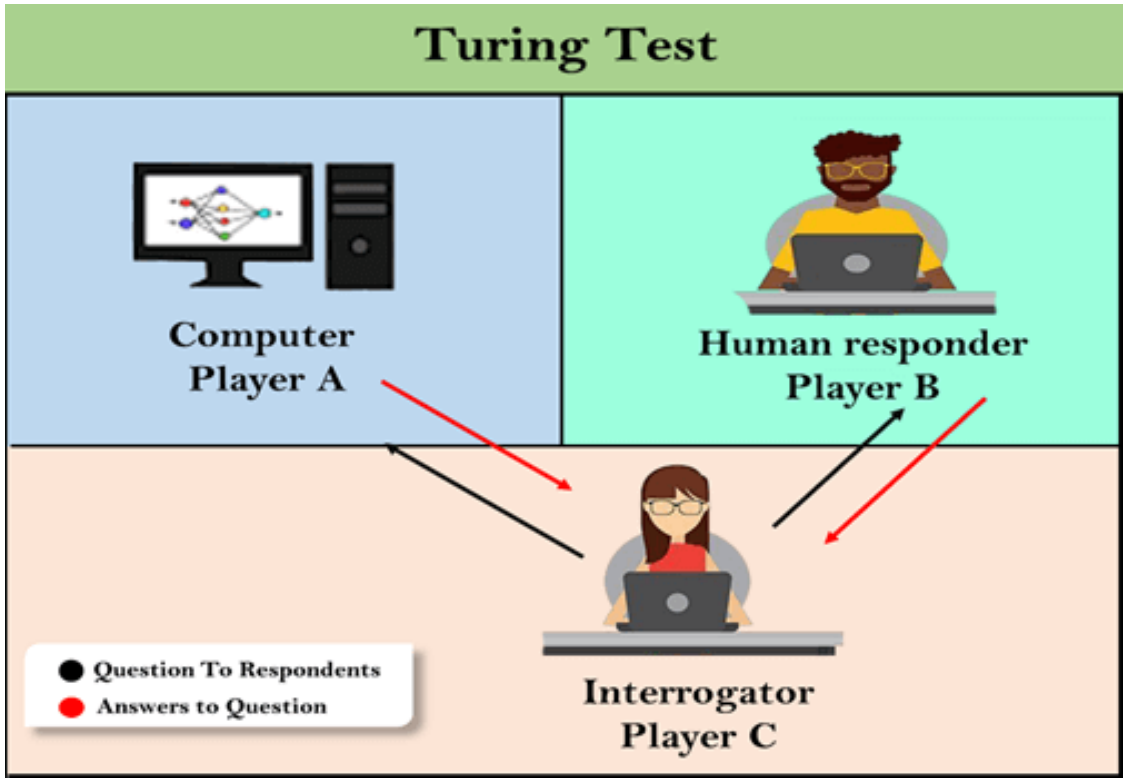


与有意识的个体（人）比较如何?

1.2 基本概念

✦图灵测试 Turing test

英国计算机科学家图灵于1950年提出的思想实验



图灵测试尚有缺陷

✦定义1-5：自然语言处理

Natural Language Processing, NLP

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）

1.2 基本概念

✦ 三种不同的语系

- **屈折语（fusional language/ inflectional language）：**

  用词的形态变化表示语法关系，如英语、法语等。

- **黏着语（agglutinative language）：**

  词内有专门表示语法意义的附加成分，词根或词干与附加成分的结合不紧密，如日语、韩语、土耳其语等。

- **孤立语（isolating language）/分析语（analytic language）：**

  形态变化少，语法关系靠词序和虚词表示，如汉语。

1.2 基本概念

- **汉语：**汉族的语言，是我国的主要语言。现代汉语的标准语是普通话。

- **中文：**中国的语言文字，特指汉族的语言文字。

—《现代汉语词典》，2016

- **定义1-6：中文信息处理**

Chinese Information Processing

针对中文的自然语言处理技术。

1.2 基本概念

中文作为使用者最多的语言，但在互联网内容上还没有成为主导性的语言。中文信息处理技术有最广泛的受众群体和最丰富的研究内容。

中文信息处理技术早已成为国际学术界和企业界共同关注的问题，汉英两大强势语言的自动翻译问题则是人类语言技术中最具挑战性的研究课题。

1.2 基本概念

✦定义1-7：人类语言技术

Human Language Technology, HLP

近几年来，自然语言处理技术迅速发展成为一门相对独立的学科，倍受关注，而且该技术不断与语音技术相互渗透和结合形成新的研究分支，因此，很多人在谈到“计算语言学”、“自然语言处理”或“自然语言理解”这些术语时，往往默认为同一个概念。甚至有专著[刘颖，2002，计算语言学]干脆直接解释为：计算语言学也称自然语言处理或自然语言理解。

1.2 基本概念

✦ 自然语言理解（Natural Language understanding, NLU）

人工智能最重要的研究方向之一，是当今“人工智能皇冠上的明珠”。

✦ 计算语言学（Computational Linguistics, CL）

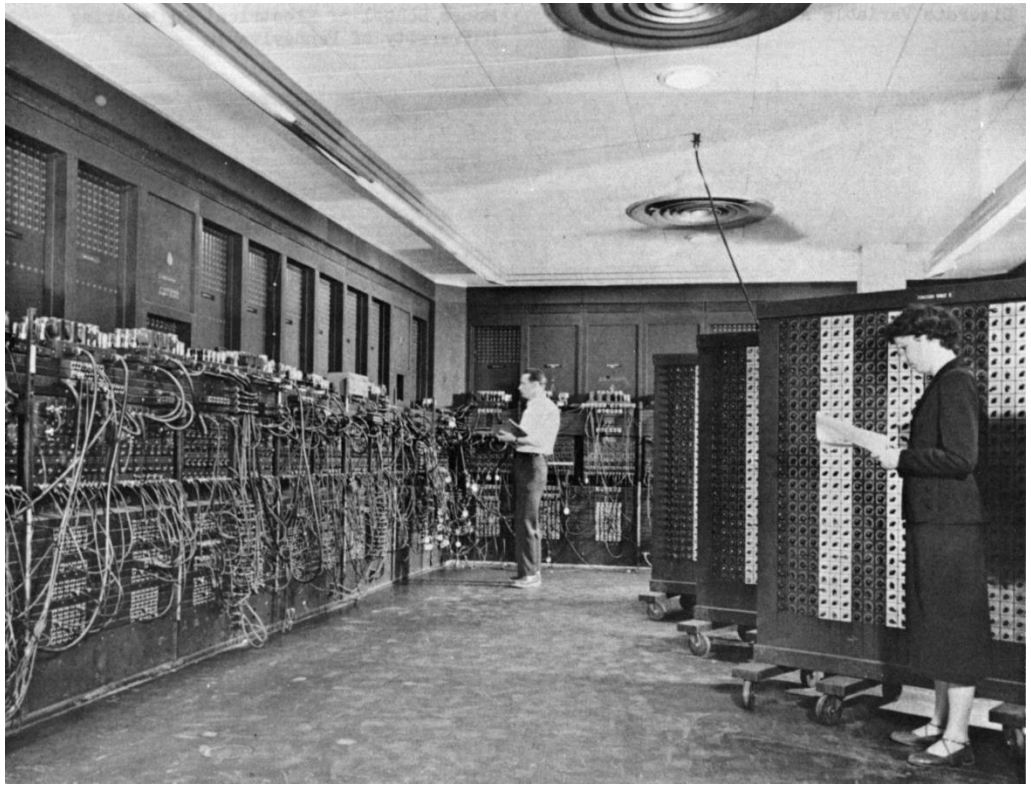
1960s，形成相对独立的学科。1962年国际计算语言学学会(ACL)成立，1965年国际计算语言学委员会(ICCL)成立，1966年“计算语言学”首次出现在美国国家科学院ALPAC报告里。

✦ 自然语言处理（Natural Language Processing, NLP）

1980S，面向计算机网络和移动通信，从系统实现和语言工程的角度开展语言信息处理方法的研究。专门针对中文的语言信 息技术研究成为中文信息处理。

1.3 NLP的产生与发展

1.3 NLP的产生与发展



1946年，世界上第一台通用计算机ENIAC诞生

Warren Weaver

- 1894.7.17 – 1978.11.24
- 信息论先驱，机器翻译早期研究者
- 1920至1932年Wisconsin大学数学教授
- 1932至1955年担任Rockefeller Institute自然科学部主任

Andrew Donald Booth

- 1918.2.11 – 2009.11.29
- 数学物理学家，二战中参与计算机研制，在程序化计算机研究中成绩卓著
- 1947年3月至9月，曾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。

1.3 NLP的产生与发展

Norbert Wiener

- 1894.11.26 – 1964.3.18
- 数学家，哲学家
- 提出控制论

[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

1.3 NLP的产生与发展

- 美国和英国的学术界对机器翻译产生了浓厚的兴趣，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
- 在随后10多年里，MT研究在国际上出现热潮，一批自然语言人机接口系统和对话系统相继出现

随着机器翻译研究的进展，各种自然语言处理技术应运而生，并逐渐发展壮大，形成了这一语言学与计算机技术相结合的新兴学科。

1.3 NLP的产生与发展

达特茅斯学院（1769建校，藤校）

左起：摩尔、麦卡锡、明斯基、 赛弗里奇(Oliver Selfridge)、所罗门诺夫

达特矛斯夏季人工智能研究计划（达特矛斯会议, 1956）

Summer Research Project on Artificial Intelligence (Dartmouth Conference)

1.3 NLP的产生与发展

- 1962年美国成立“机器翻译和计算语言学协会(Association for Machine Translation and Computational Linguistics)”并组织召开了第一届国际计算语言学学术年会(ACL)
- 1965年杂志 Machine Translation 改名为 Machine Translation and Computational Linguistics
- 1965年成立国际计算语言学委员会 (The International Committee on Computational Linguistics, ICCL)，并组织召开了第一届国际计算语言学大会 (The International Conference on Computational Linguistics, COLING)
- 1966年术语Computational Linguistics 正式出现在ALPAC (Automatic Language Processing Advisory Committee)中

1.3 NLP的产生与发展

✦ 曲折的发展历史

- 1960s 中期之前：萌芽期
- 1960s 中期到1970s 中后期：步履维艰
  - 1966年美国科学院发表 ALPAC报告
- 1970s 中后期到1980s 后期：复苏
- 1980s 至2010s 左右：快速发展
- 1980s 至2010s 左右：繁荣时期

1.4 研究内容

1.4 研究内容

- **按照应用目标划分，广义上包括：**

- **机器翻译（Machine translation, MT）**

  实现一种语言到另一种语言的自动翻译

  - **应用：** 文献翻译、网页辅助浏览
  - **代表系统：**
    - Google：https://translate.google.cn
    - 百度：https://fanyi.baidu.com/
    - Linguee：https://www.linguee.com/

1.4 研究内容

✦ 机器翻译研究现状和对机器翻译的认识

- ✦ 机器翻译研究在过去五十多年的曲折发展经历中，无论是它给人们带来的希望还是失望我们都必须客观地看到，机器翻译作为一个科学问题在被学术界不断深入研究的同时，企业家们已经从市场上获得了相应的利润。
- ✦ 在机器翻译研究中实现人机共生(man-machine symbiosis)和人机互助，比追求完全自动的高质量的翻译(Full Automatic High Quality Translation, FAHQT) 更现实、更切合实际[Hutchins, 1995]

1.4 研究内容

全球数万亿网页，
90%非汉语文字

“一带一路”沿线
65个国家和地区
50多种语言
44亿人口

1.4 研究内容

✦信息检索（Information retrieval）

信息检索也称情报检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息。

- **代表系统：**
  - 百度：https://www.baidu.com
  - Deepl：https://www.deepl.com/translator
  - Google：https://www.google.com
- 目前至少有1.8亿网站，300多亿个网页，每天数以万计地增加，只有1%的信息被有效地利用。

1.4 研究内容

✦ 自动文摘 (Automatic summarization/ Automatic abstracting)

将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。

观点挖掘 (Opinion mining)

- **应用**：电子图书管理、情报获取等。

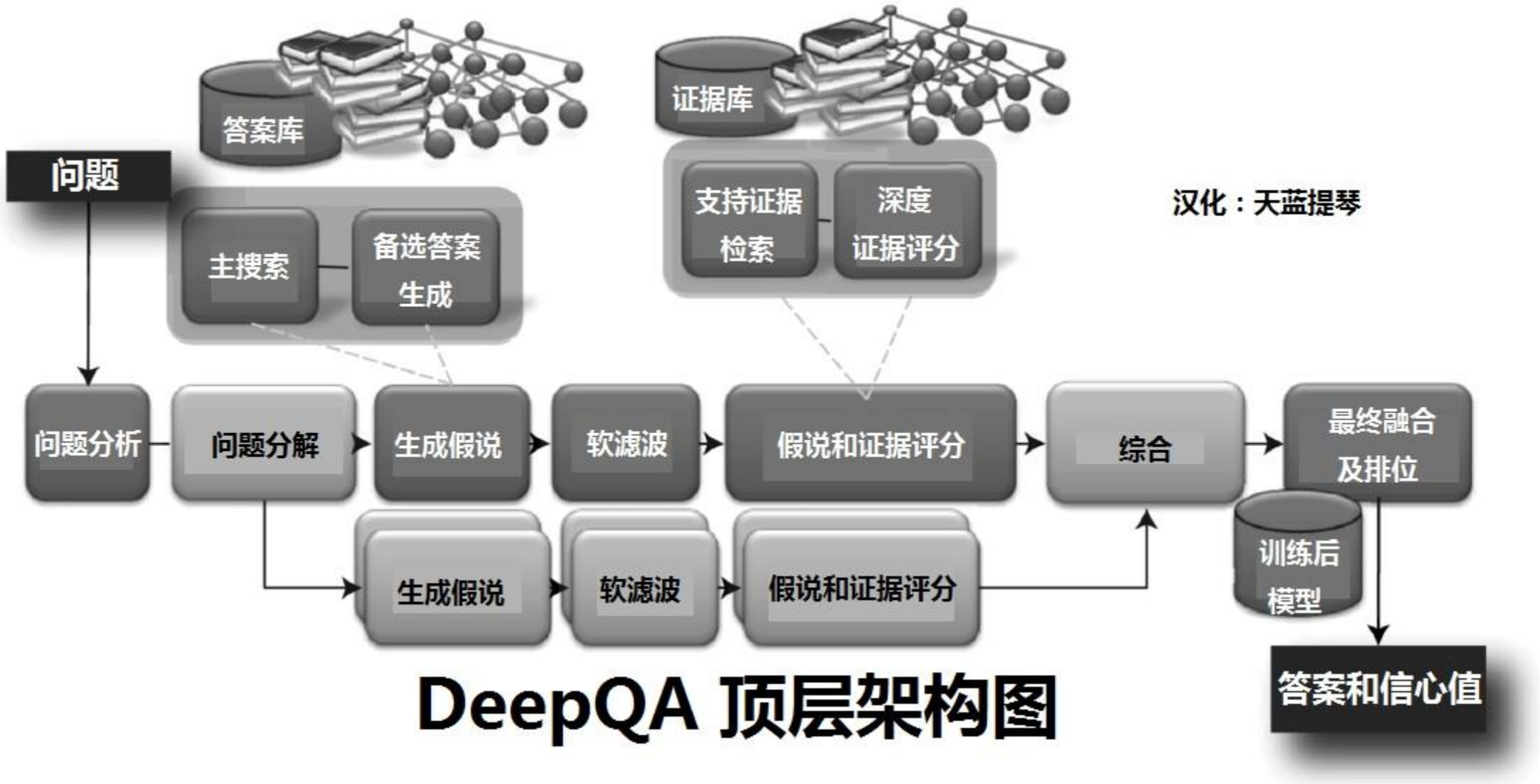
1.4 研究内容

✦ 问答系统 (Question-answering system)

通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并 做出相应的回答。问答技术有时与语音技术和多模 态输入/输出技术，以及人机交互技术等相结合，构成人机对话系统 (man-computer dialogue system)。

- **代表系统：**
  - **百度知道：**用户群体智慧
  - **IBM Watson 自动问答系统**

1.4 研究内容



DeepQA 顶层架构图

1.4 研究内容

✦信息过滤 (Information filtering)

通过计算机系统自动识别和过滤那些满足特定条件的文档信息。

✦信息抽取 (Information extraction)

从指定文档中或者海量文本中抽取出用户感兴趣的信息。

实体关系抽取 (entity relation extraction)

社会网络 (social network)

1.4 研究内容

✦文字编辑和自动校对(Automatic proofreading)

对文字拼写、用词、甚至语法、文档格式等进行自动检查、校对和编排。

- **应用：**排版、印刷和书籍编撰等。
- **代表系统：**Grammarly

✦语言教学 (Language teaching)

✦文字识别 (Character recognition)

1.4 研究内容

✦语音识别 (automatic speech recognition, ASR)

将输入语音信号自动转换成书面文字。

- **应用**：文字录入、人机通讯、语音翻译等等。
- **困难**：大量存在的同音词、近音词、集外词、口音等等。

✦文语转换/ 语音合成 (text-to-speech synthesis)

将书面文本自动转换成对应的语音表征。

- **应用**：朗读系统、人机语音接口等等。

1.4 研究内容

✦ 说话人识别/认同/验证 (speaker recognition/ identification/verification)

对一言语样品做声学分析，依此推断(确定或 验证)说话人的身份。

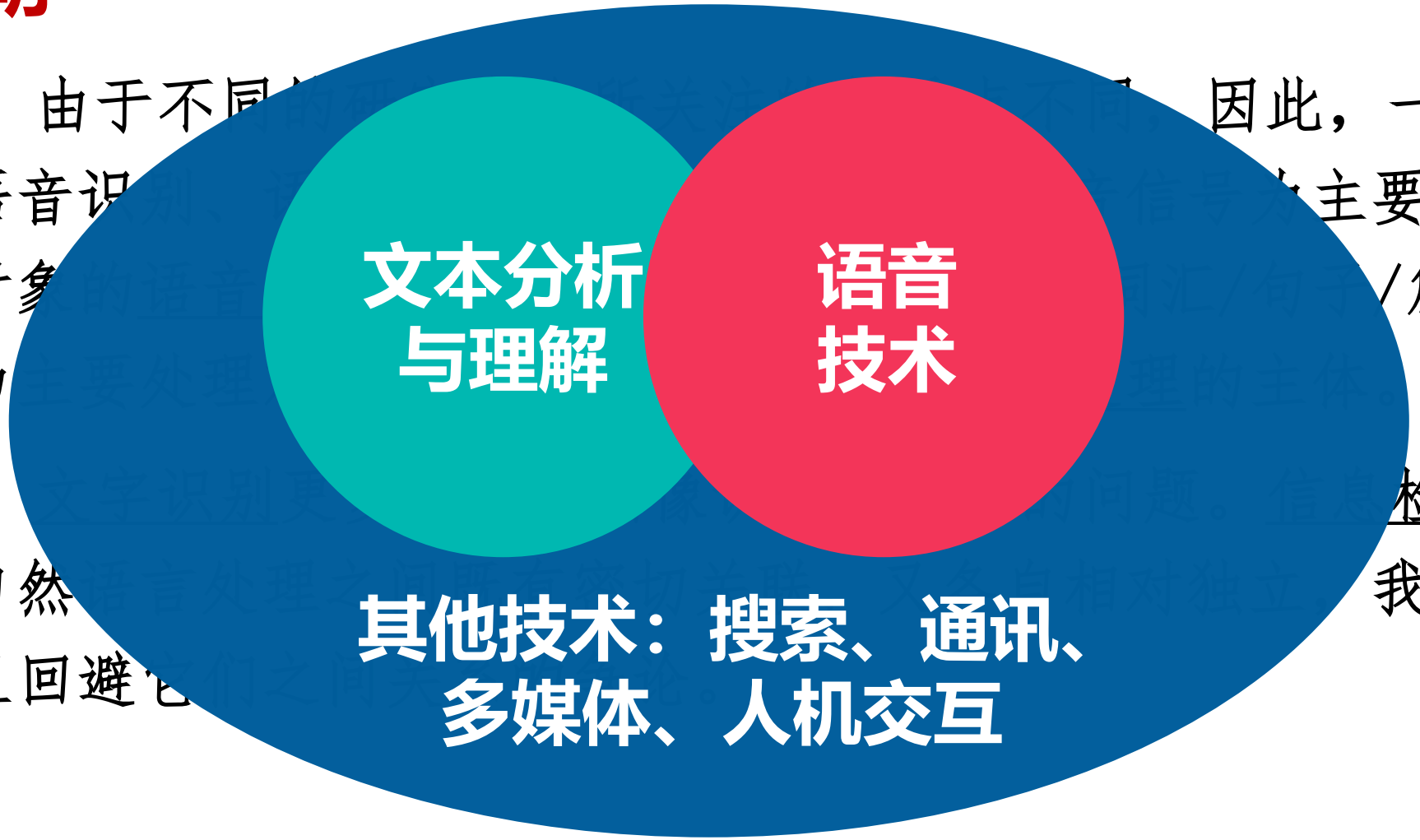
- **应用**：信息安全、防伪等等。

1.4 研究内容

✦ 说明

由于不同[illegible]因此，一般将语音识[illegible]主要研究对象的[illegible]/篇章等)为[illegible]

[illegible]检索与自然[illegible]我们暂且回避它[illegible]



1.5 基本问题和主要困难

1.5 基本问题和主要困难

✦基本问题之一：形态学(Morphology)问题

研究词(word) 由有意义的基本单位 - 词素 (morphemes) 的构成问题。

单词的识别/ 汉语的分词问题。

词素：词根、前缀、后缀、词尾

Rind 牛
Rindfleisch 牛肉
Rindfleischetikettierung 牛肉标签
Rindfleischetikettierungsüberwachung 牛肉标签监控
Rindfleischetikettierungsüberwachungsaufgaben 牛肉标签监控任务
Rindfleischetikettierungsüberwachungsaufgabenübertragung 牛肉标签监控任务转移
Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz 牛肉标签监控任务转移法

1.5 基本问题和主要困难

✦基本问题之二：句法(Syntax)问题

研究句子结构成分之间的相互关系和组成句子序列的规则 。

为什么一句话可以这么说也可以那么说?

如何建立快速有效的句子结构分析方法?

苹果，我吃了。
我吃了苹果。
≠苹果吃了我。

1.5 基本问题和主要困难

✦ 基本问题之三：语义(Semantics)问题

研究如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用来推导出该语句的意义。

这句话说了什么？

阿呆给乔局长送红包时，两人的对话颇有意思。

乔局长：“你这是什么意思？”

阿呆：“没什么，意思意思。”

乔局长：“你这就不够意思了。”

阿呆：“小意思，小意思。”

乔局长：“你这人真有意思。”

阿呆：“其实也没有别的意思。”

乔局长：“那我就不好意思了。”

阿呆：“是我不好意思。”

1.5 基本问题和主要困难

✦基本问题之四：语用(Pragmatics)问题

研究在不同上下文中语句的应用，以及上下文对语句理解所产生的影响。从狭隘的语言学观点看，语用学处理的是语言结构中有形式体现的那些语境。相反，语用学最宽泛的定义是研究语义学未能涵盖的那些意义。

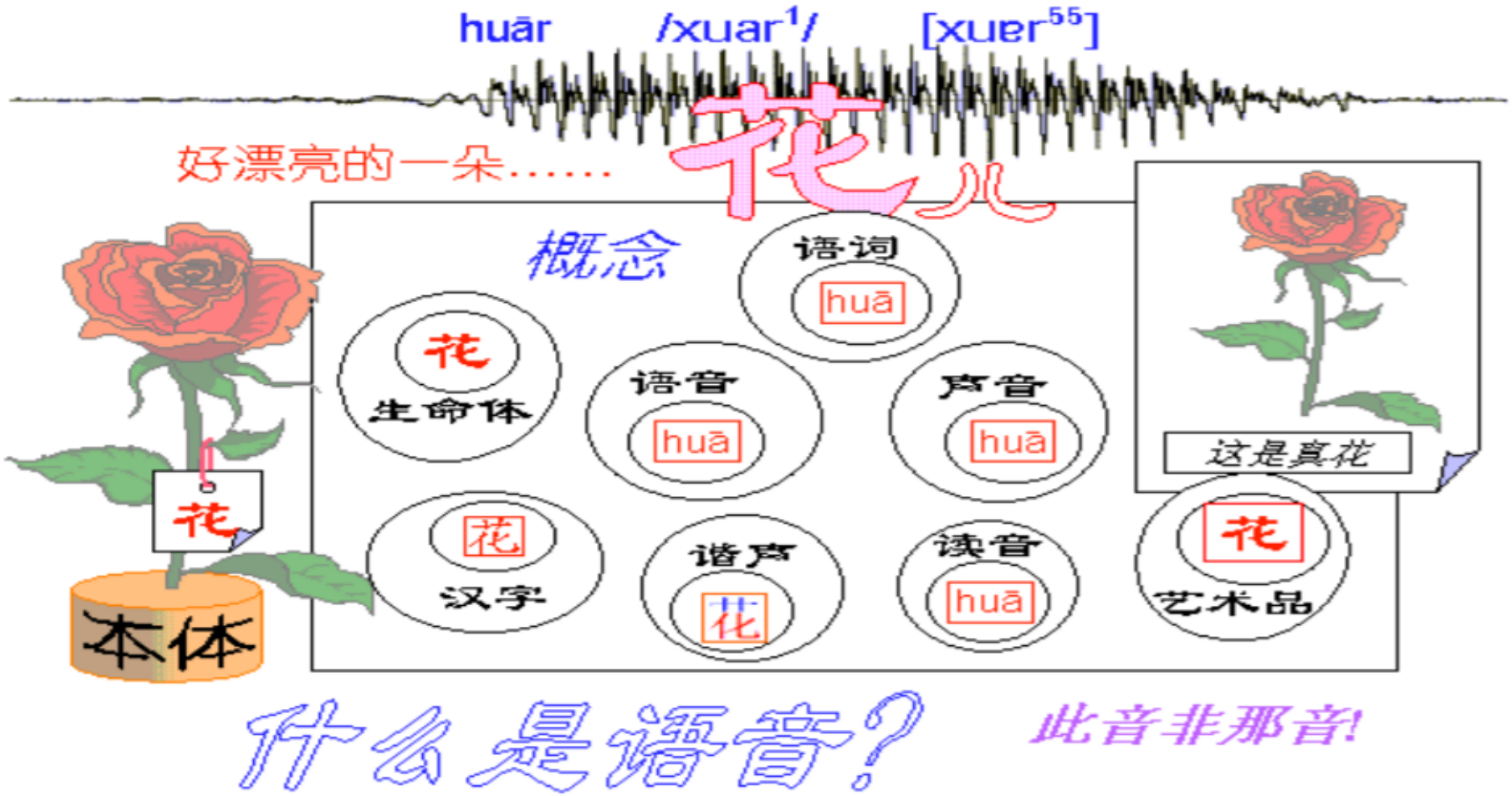
为什么要说这句话?

鲁迅：院子里有两棵树，一棵是枣树，另一棵也是枣树。

1.5 基本问题和主要困难

✦ 基本问题之五：语音学(Phonetics)问题

研究语音特性、语音描述、分类及转写方法等。

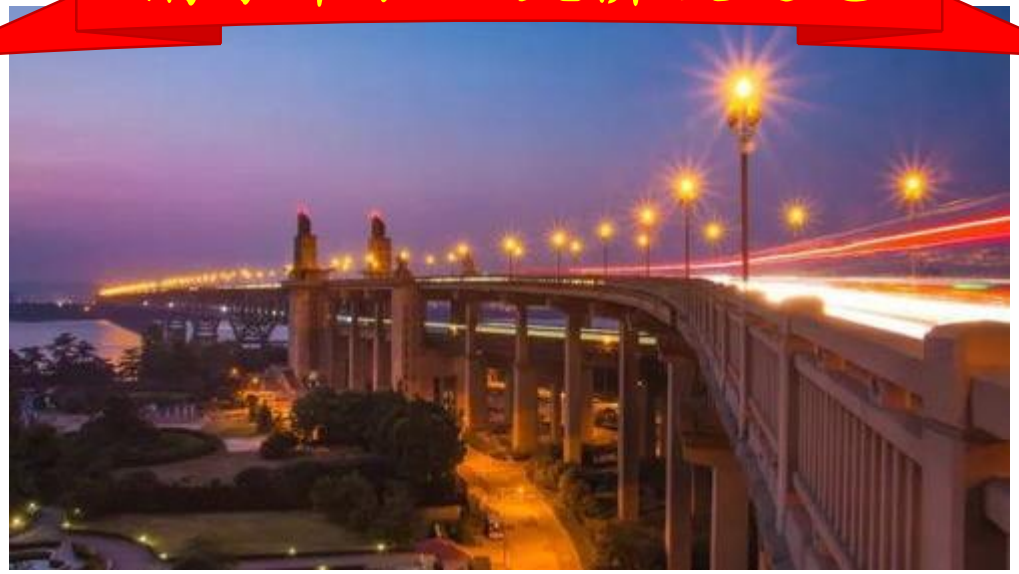


1.5 基本问题和主要困难

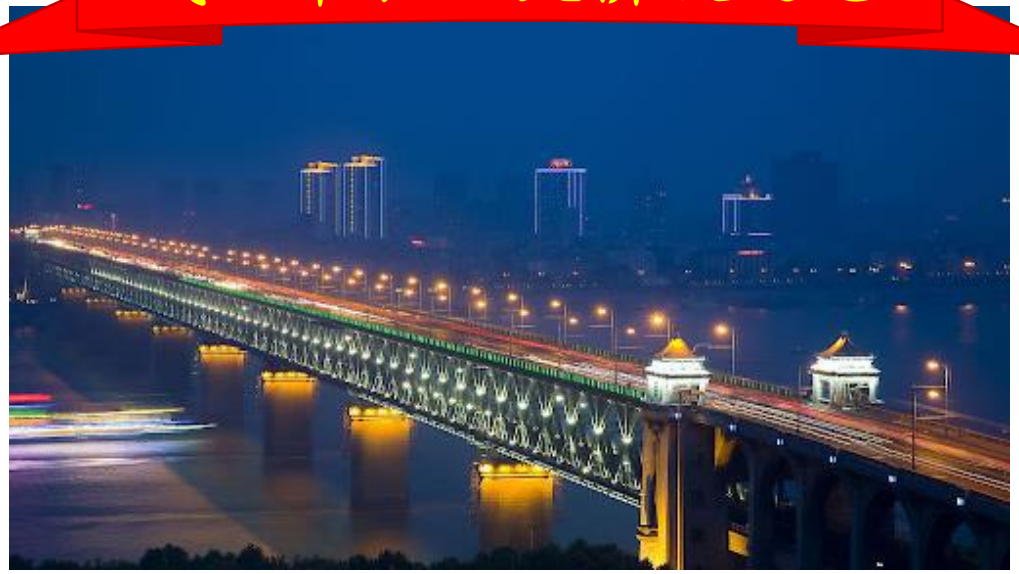
✦ 困难之一：歧义(Ambiguity)问题

➢ 词法歧义

南京市长江大桥欢迎您



武汉市长江大桥欢迎您



1.5 基本问题和主要困难

➢文章标题中的歧义

- 《汉语语法研究所面临的挑战》
  - ？ 汉语语法研究/所面临的
  - ？ 汉语语法研究所/面临的

- 《美男子整容为成“龙女”不惜割耳整鼻纹满鳞片》
  - ？ 美/男子/整容……
  - ？ 美男子/整容……

1.5 基本问题和主要困难

➢ 词性歧义

① 介词：好似，像
② 动词：喜欢

(1) Time flies like an arrow.

① 动词：飞
② 名词：苍蝇

(2) “动物保护警察”明年上岗

① 偏正短语
② 主谓宾

1.5 基本问题和主要困难

➢ 结构歧义

① 喜欢乡下的孩子。

② 关于鲁迅的文章。

③ 华强吃了西瓜。

④ 华强吃了同行。

⑤ 华强吃了官司。

⑥ 华强吃了花生米。

⑦ 写文章　　写毛笔　　写黑板

1.5 基本问题和主要困难

➢ 结构歧义

⑧ I saw a man with a telescope.

→ I saw [a man with a telescope].

→ I [saw a man] with a telescope.

? I saw a man with a telescope in the park.

英语句子歧义组合的卡特兰数(Catalan Numbers) C_n:

$$C_n = \binom{2n}{n}\frac{1}{n+1} \quad \text{其中:} \quad \binom{2n}{n} = \frac{(2n)!}{n! \times n!}$$

N为句子中介词短语的个数

1.5 基本问题和主要困难

➢ 语义歧义

➢ 他说:“她这个人真有意思(funny)”。她说:“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish),并让他向她意思意思(express)。他火了:“我根本没有那个意思(thought)”!她也生气了:“你们这么说是什么意思(intention)”?事后有人说:“真有意思(funny)”。也有人说:“真没意思(nonsense)”。

—《生活报》1994.11.13.第6版

➢ 人们的语言表达中大量地使用缩略语和隐喻的表达方式

➢ 要把权力装进制度的笼子;老虎苍蝇一起打

1.5 基本问题和主要困难

- **语音歧义**
- 初级水平

  四是四，十是十。十四是十四，四十是四十。谁要把十四说成四十就罚谁十四，谁要把四十说成十四就罚谁四十。

- 进阶水平

  石室诗士施氏，嗜狮，誓食十狮。施氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。施氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，施氏始试食是十狮尸。食时，始识是十狮尸，实十石狮尸。试释是事。

  ——《施氏食狮史》，赵元任

1.5 基本问题和主要困难

➢ 多音字及韵律等歧义

语音合成面临的诸多问题

① 一字多音

例如：小心地滑/甲壳/亲家/削铅笔

② 韵律、声调、语气、重音

例如：药材好药才好。

他的钱包被偷了。

聊吧/说吧。

1.5 基本问题和主要困难

✦ 困难之二：大量未知语言现象

- **新词、人名、地名、术语等：**

  新冠，懂王，睡王，奥力给，yyds，yygq

- **新含义：**

  小米十代，牙膏，刀法

- **新用法和新句型等：**

  工资被平均，辣眼睛，转发这条锦鲤

1.5 基本问题和主要困难

✦归纳起来，NLU 所面临的挑战：

① **普遍存在的不确定性：**词法、句法、语义、语用和语音各个层面

② **未知语言现象的不可预测性：**新的词汇、新的术语、新的语义和语法无处不在

③ **始终面临的数据不充分性：**有限的语言集合永远无法涵盖开放的语言现象

④ **语言知识表达的复杂性：**语义知识的模糊性和错综复杂的关联性难以用常规方法有效地描述，为语义计算带来了极大的困难

1.5 基本问题和主要困难

⑤ **机器翻译中映射单元的不对等性：** 词法表达不相同、句法结构不一致、语义概念不对等。

从大量复杂多样的不确定性中寻找确定性结论

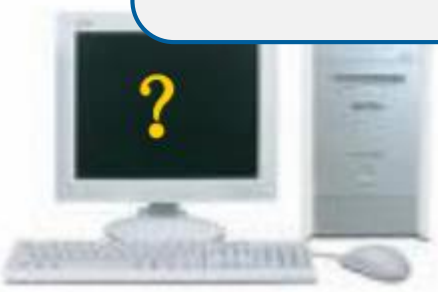
1.5 基本问题和主要困难

✦ 人脑理解语言是一个复杂的思维过程

- ❖ 语言学、心理学
- ❖ 逻辑学、认知科学
- ❖ 计算机科学
- ❖ 统计学、信息论
- ❖ 背景知识、常识
- ❖ ……

1.5 基本问题和主要困难

人脑的语言认知过程到底怎样?



文章讲了什么?

1.6 基本研究方法

1.6 基本研究方法

✦ 理性主义（Rationalism）

通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识，而这些语句和语言现象在实际的应用中并不常见。

- **理论基础：Chomsky的文法理论**
- **问题求解的基本思路：基于规则的分析方法建立符号处理系统**
- **规则库开发：** N + N → NP
- **词典标注：** #工作, N(uc); V;
- **推导算法设计：** 归约、推导、歧义消解方法…

知识库 + 推理系统 → NLP系统

1.6 基本研究方法

✦ 经验主义 (Empiricism)

偏重于对大规模语言数据中人们所实际使用的普通语句的统计。

- **理论基础：统计学、信息论、机器学习**
- **问题求解的基本思路：基于大规模真实语料（语言数据）建立计算方法**
- **大规模真实数据的收集、标注：**真实性、代表性、标注信息...
- **统计模型建立：**模型的复杂性、有效性、参数训练方法 .....

语料库 + 统计模型 → NLP系统

1.6 基本研究方法

- **以机器翻译为例**

- **将英语句子翻译成汉语：**

- **There is a book on the desk.**

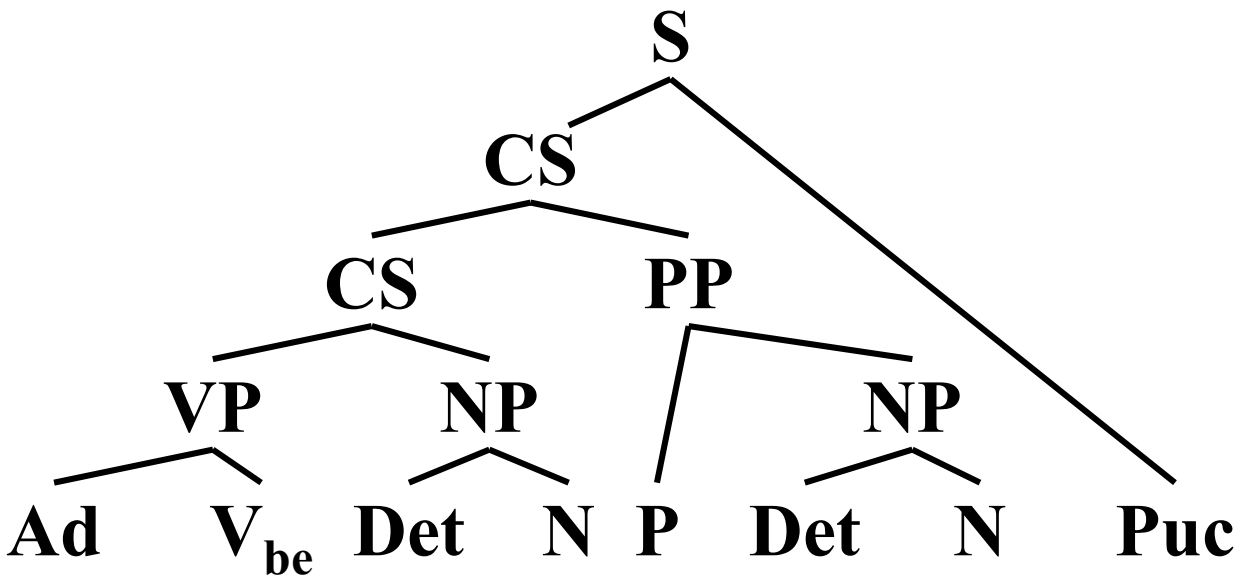
1.6 基本研究方法

✦ 基于规则的方法

① 对英语句子进行词法分析

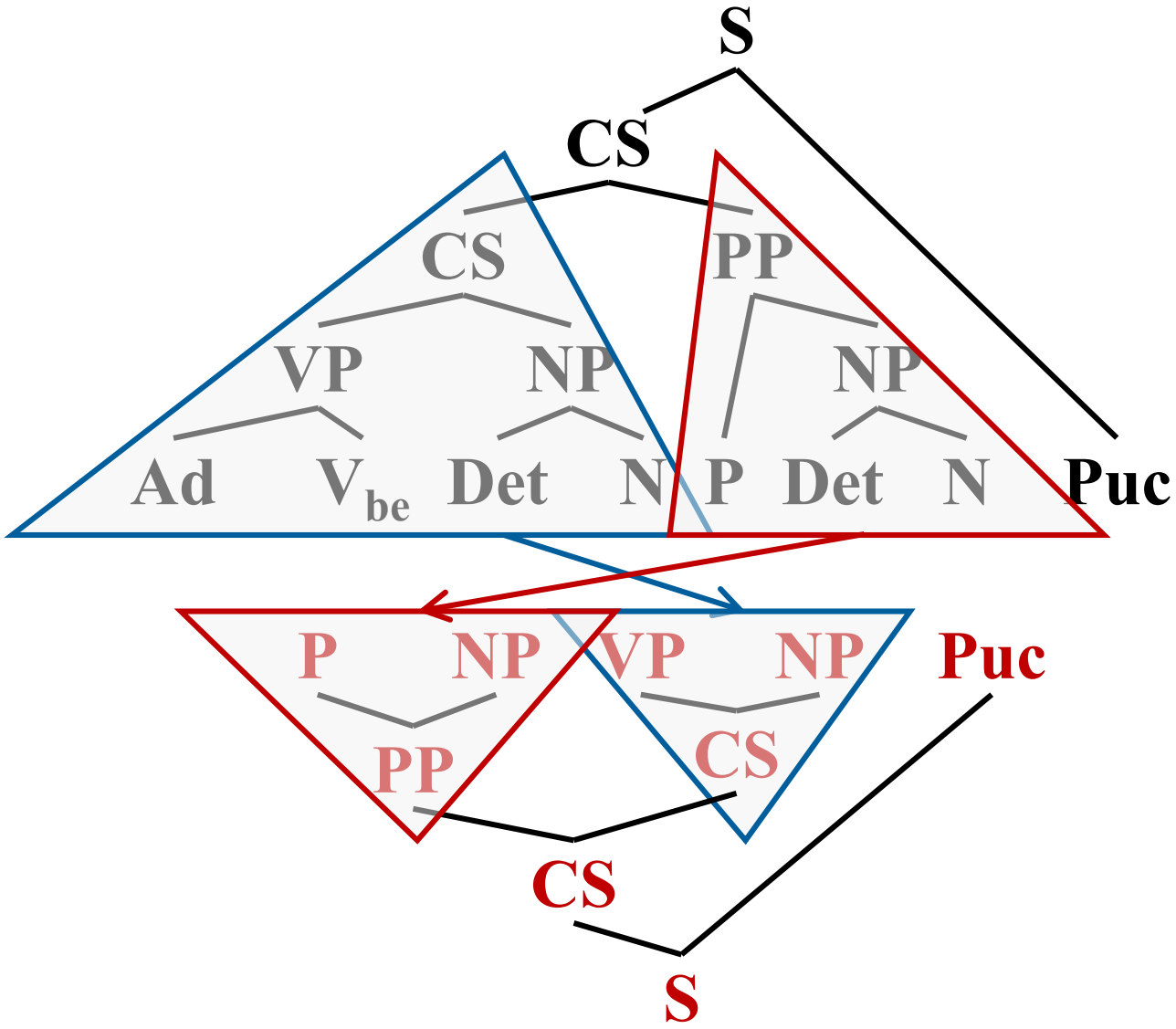
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

② 对英语句子进行句法结构分析



1.6 基本研究方法

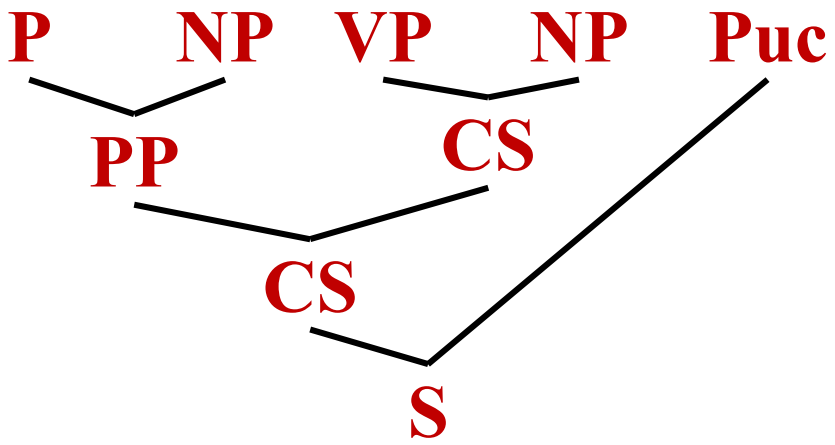
③ 利用转换规则将英语句子结构转换成汉语句子结构



1.6 基本研究方法

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

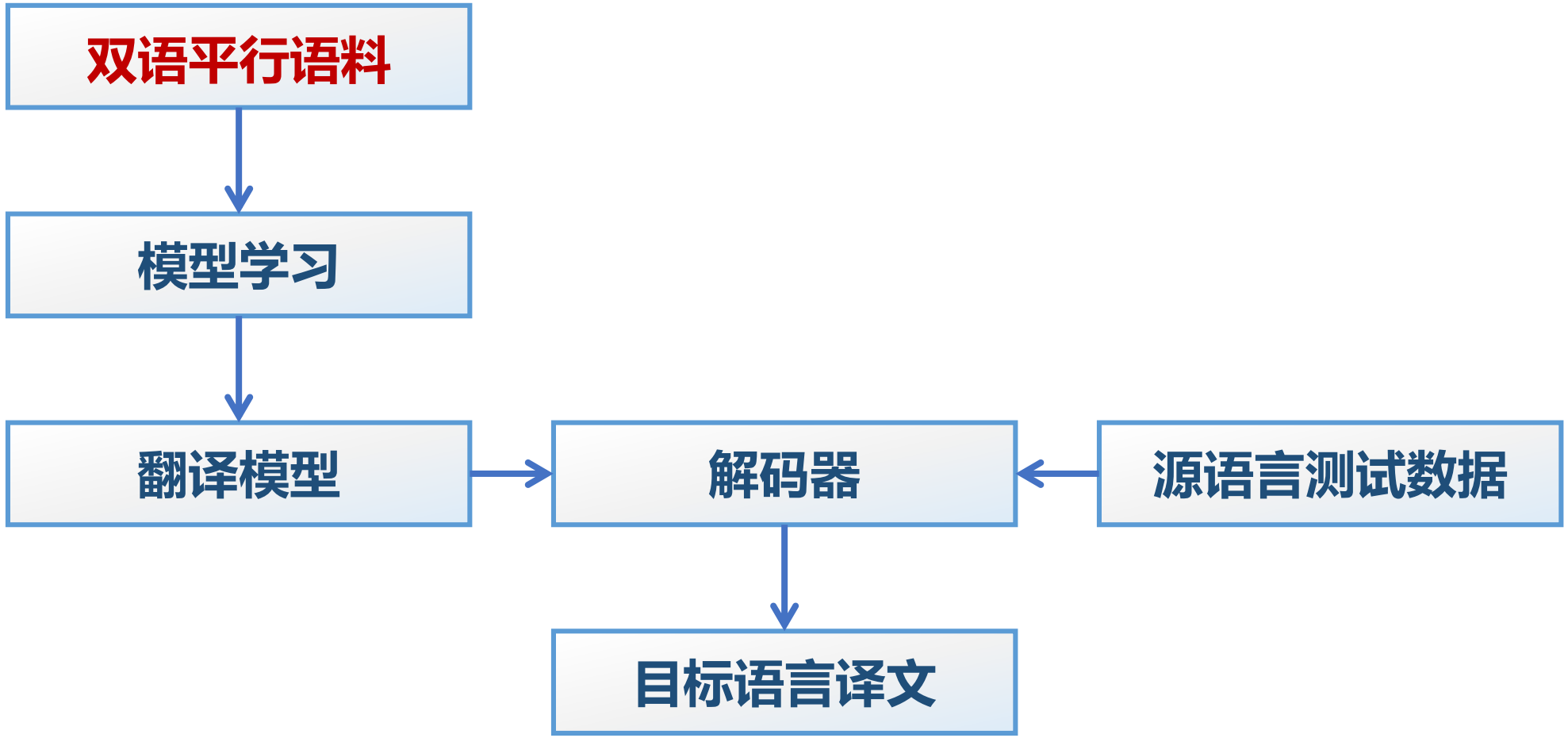
#a, Det, 一
#book, N, 书; V, 预订
#desk, N, 桌子
#on, P, 在 X 上
#There be, V, 有



中文译文：在桌子上有一本书

1.6 基本研究方法

✦ 数据驱动的翻译方法（如SMT和NMT）



1.6 基本研究方法

- hazir men gowuyüenge wakaliten yighingha hökümet xizmitidin doklat bërimen, qarap ciqishinglarni hemde memliketlik siyasiy këngesh ezalirining pikir bërishini soraymen.
- 现在，我代表国务院，向大会报告政府工作，请予审议，并请全国政协委员提出意见。
- ötken bir yil — partiye 19-qurultiyining rohini omumyüzlük izcillashturush bashlanghan yil, mushu nöwetlik hökümet qanun boyice wezipe öteshke bashlighan tunji yil.
- 过去一年是全面贯彻党的十九大精神开局之年，是本届政府依法履职第一年。
- dölitimiz tereqqiyatta köp yillardin bëri kem körülgen murekkep, keskin icki-tashqi weziyetke duc keldi, iqtisadta yëngidin töwenlep këtish bësimi körüldi.
- 我国发展面临多年少有的国内外复杂严峻形势，经济出现新的下行压力。

1.6 基本研究方法

✦ 基于统计的方法

- **给定源语言句子：** $E = e_1^m = e_1 e_2 \cdots e_m$
- **将其翻译成目标语言句子：** $C = c_1^l = c_1 c_2 \cdots c_l$
- **跟进贝叶斯公式：** $P(C|E) = \dfrac{P(C)P(E|C)}{P(E)}$

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, P(C)P(E|C)$$

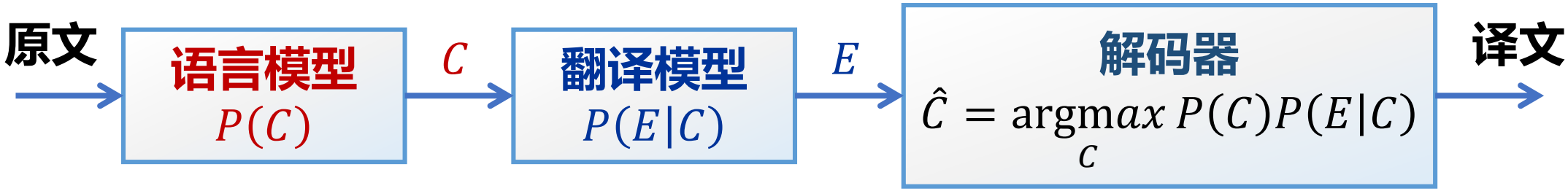
求解使P值最大的C

语言模型
Language model, LM

翻译模型
Translation model, TM

1.6 基本研究方法

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)×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

1.6 基本研究方法

规则方法

- 形式语言
- 语法理论
- 词法理论
- 推理方法
- ……

统计方法

- 数学模型
- 信息论
- 机器学习
- 搜索算法
- ……

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

1.7 研究现状

1.7 研究现状

✦ 基本现状

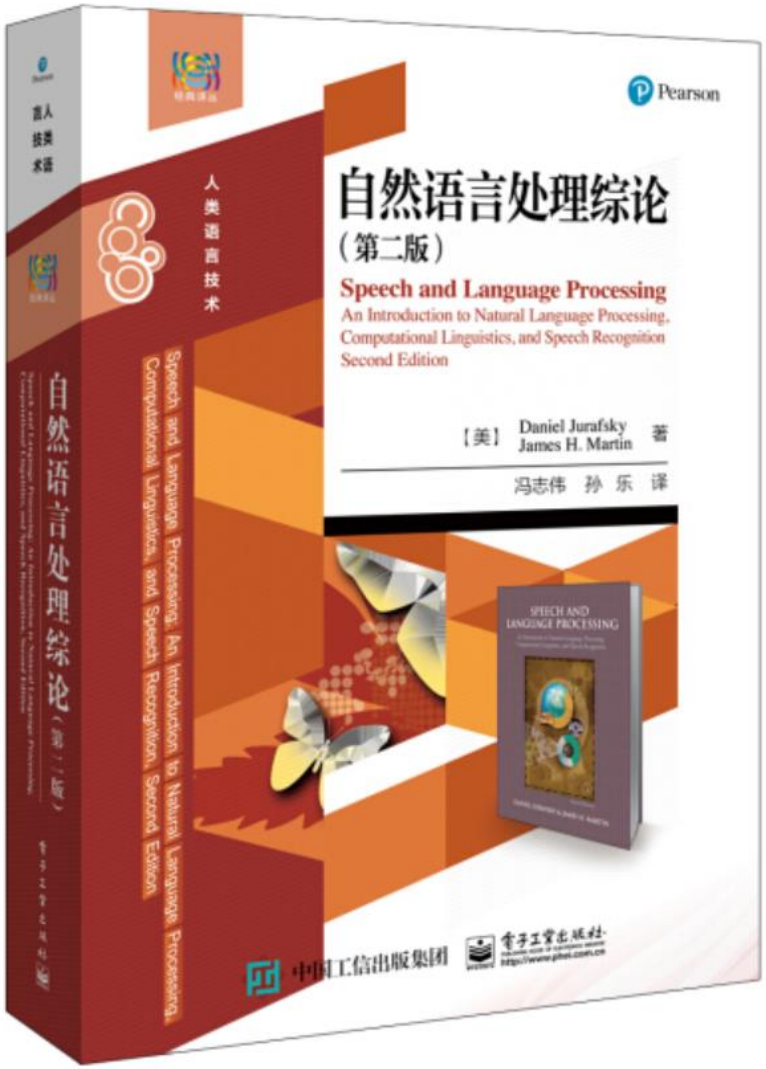
- 部分问题得到了解决，可以为人们提供辅助性 帮助，如：专业领域文档翻译，电子词典，搜 索引擎，文字录入等；
- [illegible]算、
- 社会需求日益迫切：信息服务、通讯、网络内 容管理、情报处理、国家安全等；
- 许多技术离真正实用的目标还有相当的距离， 尚未建立起有效、完善的理论体系。

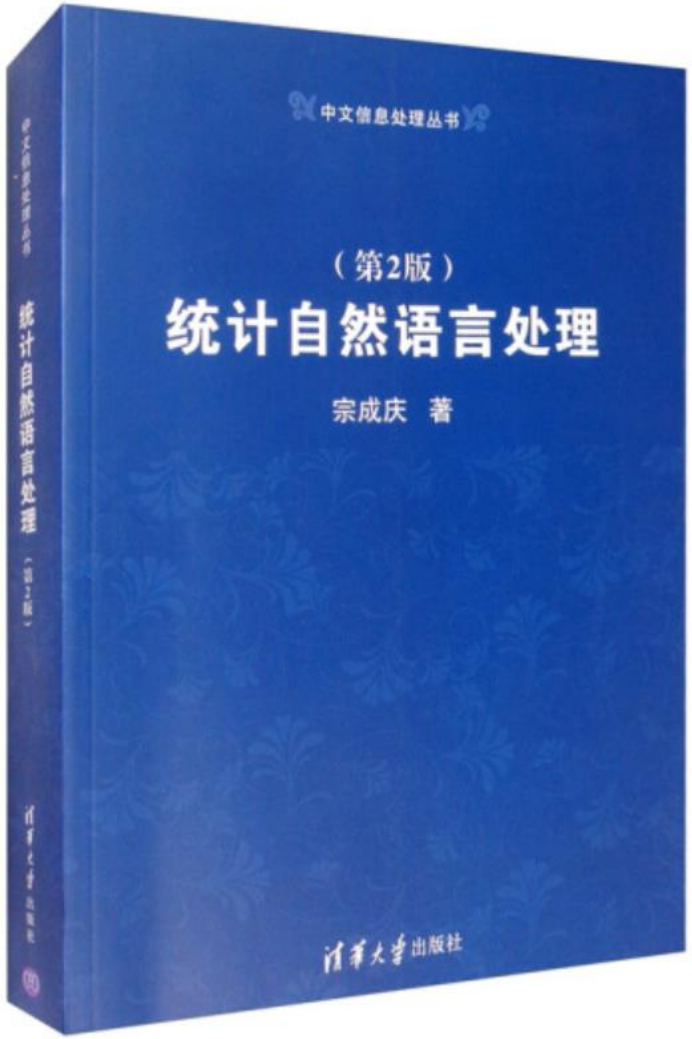
革命尚未成功，同志仍需努力！

1.8 参考文献

1.8 参考文献

✦ 专著


人类语言技术
自然语言处理综论
（第二版）
Speech and Language Processing
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
【美】 Daniel Jurafsky James H. Martin 著
冯志伟 孙乐 译
Pearson
中国工信出版集团


中文信息处理丛书
（第2版）
统计自然语言处理
宗成庆 著
清华大学出版社

1.8 参考文献

✦期刊

1. Computational Linguistics
2. Natural Language Engineering
3. ACM TALLIP
4. Machine Translation
5. IEEE Trans. on Audio, Speech, and Language Processing
6. 中文信息学报/ 计算机学报/ 软件学报/ 计算机研究与发展

1.8 参考文献

✦会议论文集

1. Proceedings of ACL (Annual Meeting of the Association for Computational Linguistics)

2. Proceedings of NAACL, EMNLP

3. Proceedings of COLING (International Conference on Computational Linguistics)

4. Proceedings of IJCNLP (International Joint Conference on Natural Language Processing)

5. 国内相关会议论文集

本章小结

- 基本概念
- 产生与发展
- 研究内容：机器翻译、信息检索...
- 基本问题：从词法、句法、语义到语用、语音
- 困难与挑战：歧义、未知现象 ...
- 研究方法：经验主义方法与理性主义方法
- 参考文献

谢 谢