数据科学与计算机学院
School of Data and Computer Science

自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

展示（一）

展示（二）

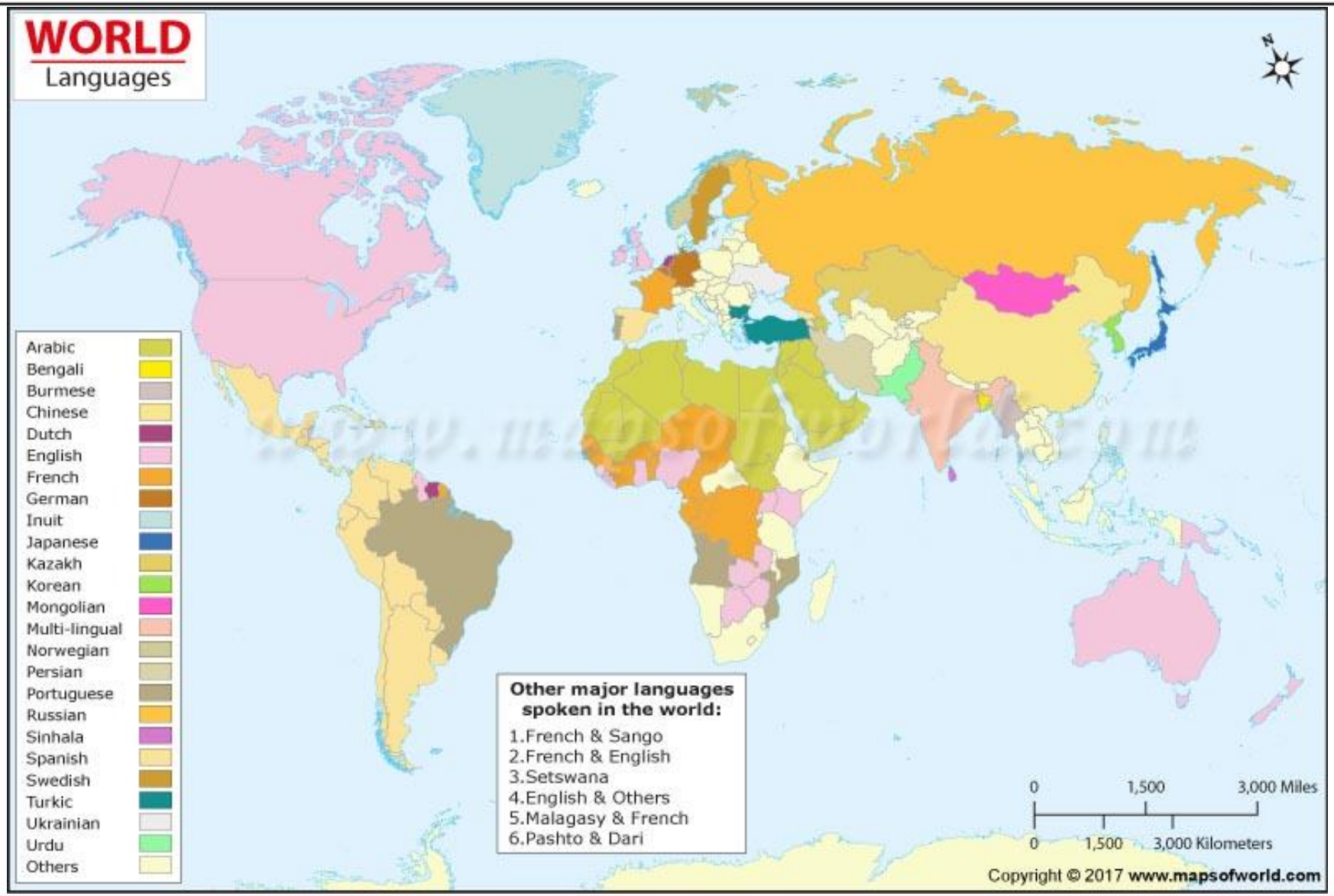
展示（三）

机器翻译概论

人类为什么会有那么多不同的语言?

巴别塔 (Tower of Babel)

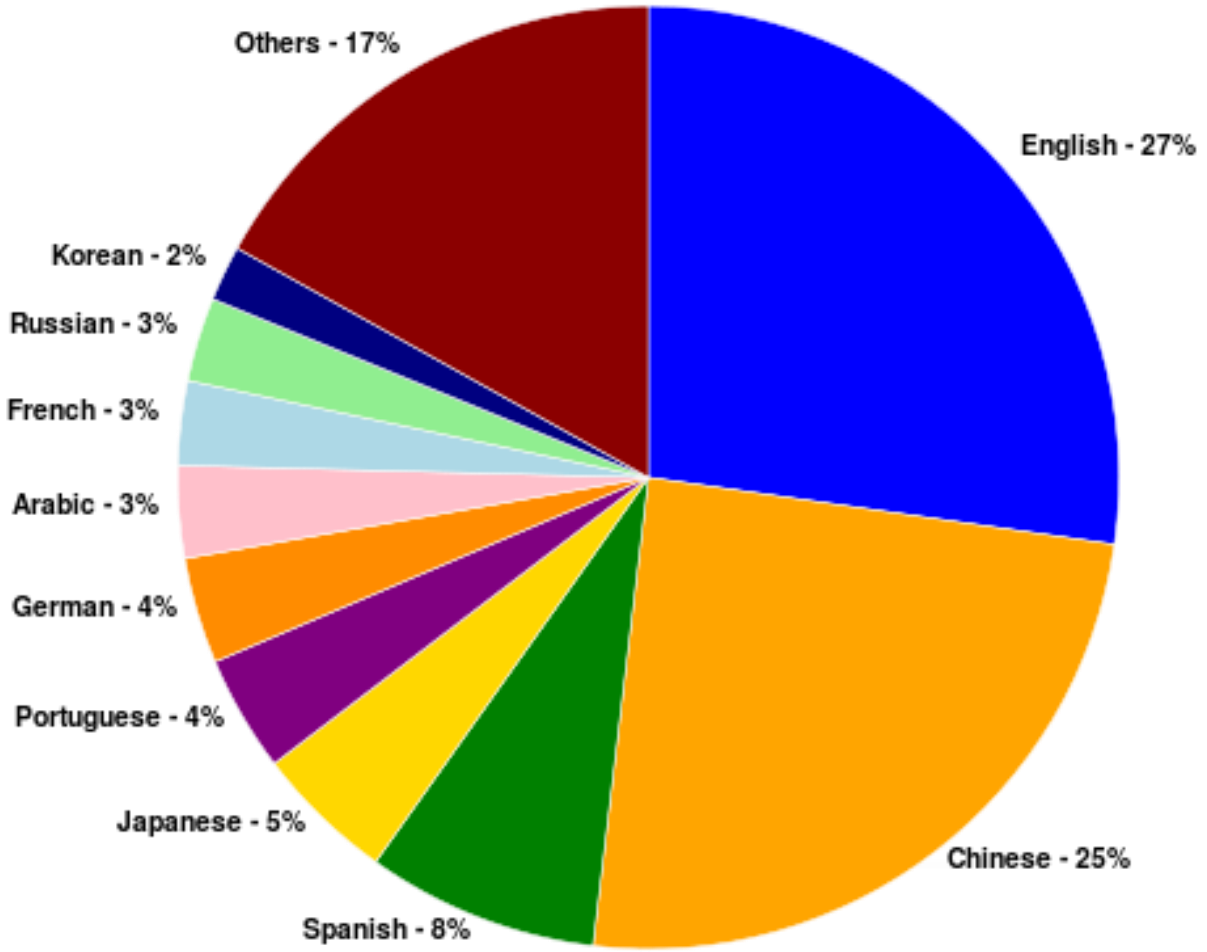
世界语言地图


WORLD
Languages
Arabic
Bengali
Burmese
Chinese
Dutch
English
French
German
Inuit
Japanese
Kazakh
Korean
Mongolian
Multi-lingual
Norwegian
Persian
Portuguese
Russian
Sinhala
Spanish
Swedish
Turkic
Ukrainian
Urdu
Others
Other major languages spoken in the world:
1.French & Sango
2.French & English
3.Setswana
4.English & Others
5.Malagasy & French
6.Pashto & Dari
0 1,500 3,000 Miles
0 1,500 3,000 Kilometers
Copyright © 2017 www.mapsofworld.com

中国语言地图


Oroqen
Evenki
Buriat
Daur
Mandarin
Manchu
Daur
Kazakh
Xibe
Korean
Mongolian
Kyrgyz
Kalmyk
Manchu
Uygur
Korean
Manchu
Mandarin
Wakhi
Yugur
Jinyu
Kalmyk
Tu
Salar
Amdo
Mandarin
Choni
Baima
Tibetan
Panang
Qiang
Jiarong
Wu
Huizhou
Sherpa
Lhomi
Groma
Luoba
Drung
Nuosu
Hmong
Miao
Xiang
Gan
Min-Bei
Min-Dong
Lisu
Naxi
Lipo
Miao
Pu-Xian
Xishanba
Bai
Nasu
Bouyei
Dong
Hakka
Min-Zhong
Lolopo
Axi
Tai-Nua
Zhuang
Yue
(Cantonese)
Min-Nan
Nisu
Riang
Parauk
Hani
Ping
Lu
Bugan
Vo-Wa
Honi
Lahu
Biyo
Min-Nan
Hlai

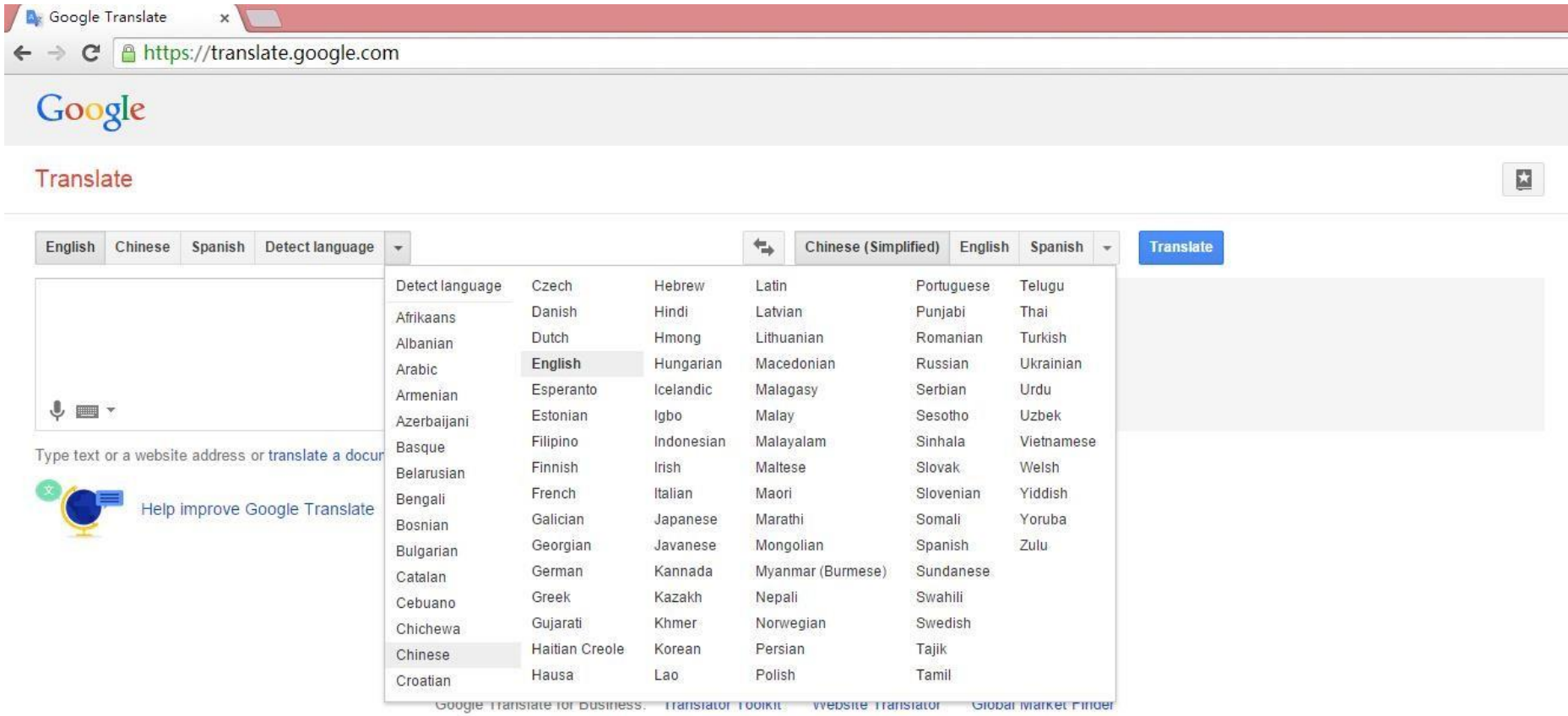
互联网用户


Others - 17%
English - 27%
Korean - 2%
Russian - 3%
French - 3%
Arabic - 3%
German - 4%
Portuguese - 4%
Japanese - 5%
Spanish - 8%
Chinese - 25%

引言

- 有关专家已经指出，语言障碍是21世纪国际社会全球化面临的主要困难之一;
- 机器翻译涉及语言学、计算语言学、认知科学和数学等多种学科，具有重要的科学意义;
- 机器翻译具有巨大的社会需求，不仅具有极大的经济利益，而且对于情报获取、信息安全意义重大;
- 以欧洲为例，有380多种语言，2004年5月1日以前欧盟有11种官方语言，每年为这11种语言翻译、转录文件耗费的人力费用大约549M欧元。

引言


Google Translate
https://translate.google.com
Google
Translate
English Chinese Spanish Detect language
Chinese (Simplified) English Spanish
Translate
Type text or a website address or translate a docu
Help improve Google Translate
Detect language
Afrikaans
Albanian
Arabic
Armenian
Azerbaijani
Basque
Belarusian
Bengali
Bosnian
Bulgarian
Catalan
Cebuano
Chichewa
Chinese
Croatian
Czech
Danish
Dutch
English
Esperanto
Estonian
Filipino
Finnish
French
Galician
Georgian
German
Greek
Gujarati
Haitian Creole
Hausa
Hebrew
Hindi
Hmong
Hungarian
Icelandic
Igbo
Indonesian
Irish
Italian
Japanese
Javanese
Kannada
Kazakh
Khmer
Korean
Lao
Latin
Latvian
Lithuanian
Macedonian
Malagasy
Malay
Malayalam
Maltese
Maori
Marathi
Mongolian
Myanmar (Burmese)
Nepali
Norwegian
Persian
Polish
Portuguese
Punjabi
Romanian
Russian
Serbian
Sesotho
Sinhala
Slovak
Slovenian
Somali
Spanish
Sundanese
Swahili
Swedish
Tajik
Tamil
Telugu
Thai
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
Welsh
Yiddish
Yoruba
Zulu

103 种语言

引言


Bai du 翻译
AI同传
自动检测
中文
翻 译
人工翻译
自动检测
英语
中文
法语
-常用语种
ABC
阿拉伯语
爱沙尼亚语
保加利亚语
波兰语
DEFG
丹麦语
德语
俄语
法语
芬兰语
HIJKLMN
韩语
荷兰语
捷克语
罗马尼亚语
OPQRST
葡萄牙语
日语
瑞典语
斯洛文尼亚语
泰语
UVWX
文言文
西班牙语
希腊语
匈牙利语
YZ
中文
英语
意大利语
越南语
粤语
中文繁体

28种语言

机器翻译的产生与发展

机器翻译的产生与发展

- 概念：机器翻译 (machine translation, MT) 是一门用计算机把一种语言(源语言，source language) 翻译成另一种语言(目标语言，target language) 的学科和技术

机器翻译的产生与发展

- **机器翻译的概念最早可追溯到17世纪**

  - 1629年，哲学家笛卡儿提出了世界语言的概念，即将不同语言相同含义的词汇以统一符号表示;
  - 笛卡儿、莱布尼兹、贝克、基尔施以及贝希尔等人曾试图编写类似世界语言的辞典;

机器翻译的产生与发展

❑ 20世纪初期:提出机器翻译的理论与想法

- 沃伦·韦弗被誉为机器翻译的鼻祖。他抛却了俄语文本的含义，转而视为一堆“密码”；
- 在美国和欧洲，他的团队和继任者在工作时都遵循着一个常理：“任何语言都是由一堆词汇和一套语法规则组成。只要把两种词汇放到机器里，按照人类组合这两种词汇的方式，为之建立一套完整的规则，机器就能破译“密码”。

机器翻译的产生与发展

翻译类似于解读密码的过程：当我阅读一篇用俄语写的文章时，我可以说这篇文章实际上是用英文写的，只不过它用另外一种奇怪的符号编了码，当我阅读时，我是在进行解码。

--1949

机器翻译的产生与发展

❑ 1954年：机器翻译可行的开端

- 1954年美国乔治城大学在一项实验中，成功将约60句的俄文自动翻译成英文，被视为机器翻译可行的开端。
- 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。

机器翻译的产生与发展

❑ 1966年：进入低潮

- 自动语言处理顾问委员会（ALPAC）在1966年提出的一项报告中表明十年来的机器翻译研究进度缓慢，未达预期。该项报告使得之后的研究资金大为减缩；

机器翻译的产生与发展

❑ 1980年代：复苏

- 由于电脑运算科技的进步，以及演算成本相对降低，才使政府与企业对机器翻译再次提起兴趣，特别是在统计法机器翻译的领域上；

机器翻译的产生与发展

❑ 1990～1999年

- 统计机器翻译由于仅依赖双语平行语料，大大降低了进入机器翻译研究的门槛；但因为数学模型过于复杂，很难实现。

机器翻译的产生与发展

❑ 2014~现在：突破

- 基于深度学习的统计机器翻译 （Devlin et al., 2014）
- 神经网络机器翻译（Sutskever et al., 2014, Bahdanau et al., 2014）

机器翻译的困难

机器翻译的困难

- 自然语言中普遍存在的歧义和未知现象
  - 句法结构歧义/词汇歧义/语用歧义…
  - 新的词汇、术语、结构、语义 …

机器翻译的困难

- 机器翻译不仅仅是字符串的转换
  - 不同语言之间文化的差异
  - 现有方法无法表示和利用世界知识和常识

机器翻译的困难

- 机器翻译的解不唯一，而且始终存在的人为的标准

自然语言处理中的很多问题在机器翻译中都会遇到

机器翻译的困难

《望庐山瀑布》

日照香炉生紫烟，

遥看瀑布挂前川。

飞流直下三千尺，

疑是银河落九天。

机器翻译的困难

《望庐山瀑布》

日照香炉生紫烟，

遥看瀑布挂前川。

飞流直下三千尺，

疑是银河落九天。

"Looking at the waterfall on Mount Lu"

Sunshine censers produce purple smoke,

Look at the waterfall hanging from QianChuan.

Three thousand feet down,

It is suspected that the Milky way has set for nine days.

百度翻译

机器翻译的困难

机器翻译的困难

机器翻译研究现状

机器翻译研究现状

- **若干翻译系统已实用化或接近实用化**
  - Google translator
  - 百度在线翻译
- **翻译质量在稳步提升**
  - 模型
  - 数据
  - 工程

机器翻译研究现状

❑ 新闻翻译

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

Chinadaily, 8 May 2012

机器翻译研究现状

- 北京做第三严正交涉到马尼拉，并警告说这是很难约领土僵局的一个岛屿乐观。当局说，他们已经准备了马尼拉的情况 有任何升级。(2015.4.28)
- 北京由第三严正交涉到马尼拉，并警告说这是很难约了一个岛领土僵局持乐观态度。当局说，他们已经为马尼拉局势的 升级准备。(2016.5.1)
- 北京对马尼拉进行了第三次庄严的代表，并警告说，对岛上的领土僵局很难看好。当局表示，他们为马尼拉的情况升级 做好了准备。(2017.4.16)
- 北京向马尼拉提出了第三次严正交涉，并警告说很难对一个岛屿的领土僵局持乐观态度。当局表示，他们已经为马尼拉局势升级做好了准备。 (2018.5.15)

机器翻译研究现状

❑ 基本观点

- 我们需要的是计算机帮助人类完成某些翻译工作，而不是完全替代人，人与机器翻译系统之间应该是互补的关系，而不是相互竞争 [Hutchins, 2001].
- “信、达、雅”是人类翻译追求的目标，计算机在这方面很难替代人.

基本翻译方法

基本翻译方法

- 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

基本翻译方法(一)：直接转换法

❑ 直接转换法

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。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. → Me(I) gusta(like) Maria(Mary).

基本翻译方法(二)：基于规则

❑ 基于规则的翻译方法(Rule-based)

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。

基本翻译方法(二)：基于规则

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

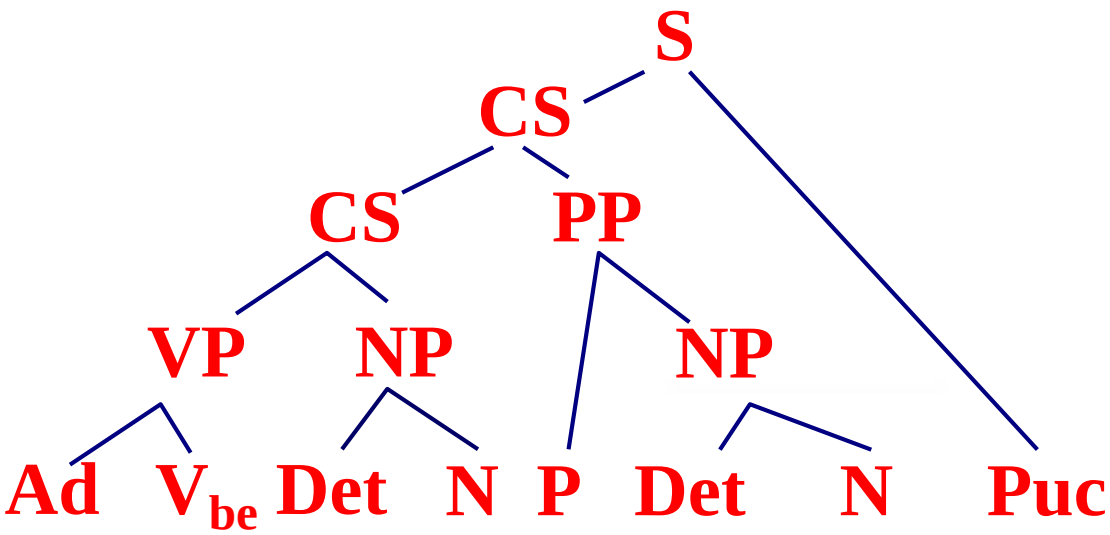
基本翻译方法(二)：基于规则

给定源语言句子： There is a book on the desk.

■Step 1: 词法分析：

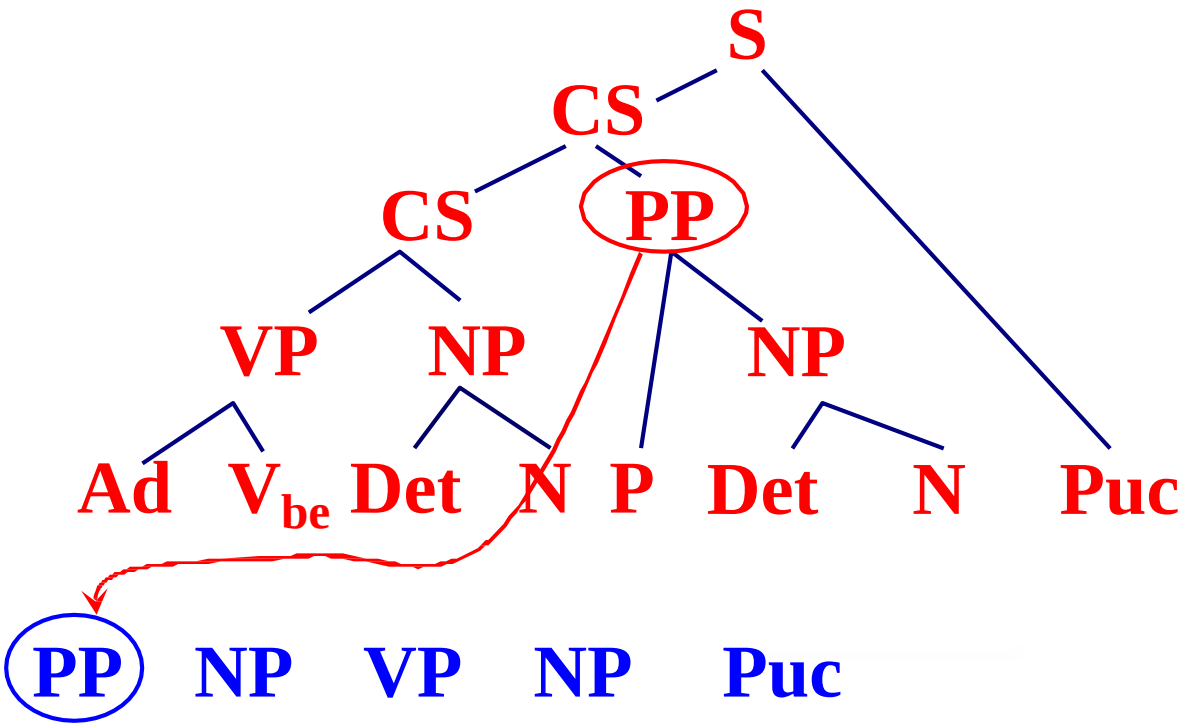
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

■Step 2: 利用句法规则进行句法结构分析：



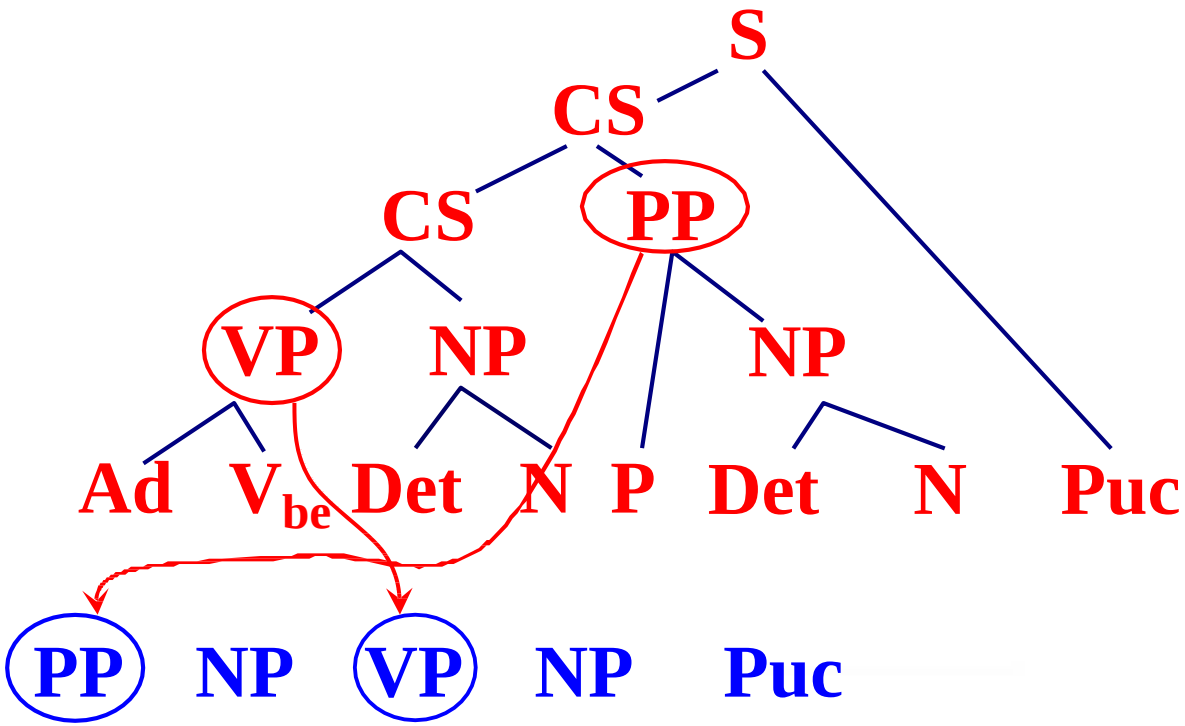
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



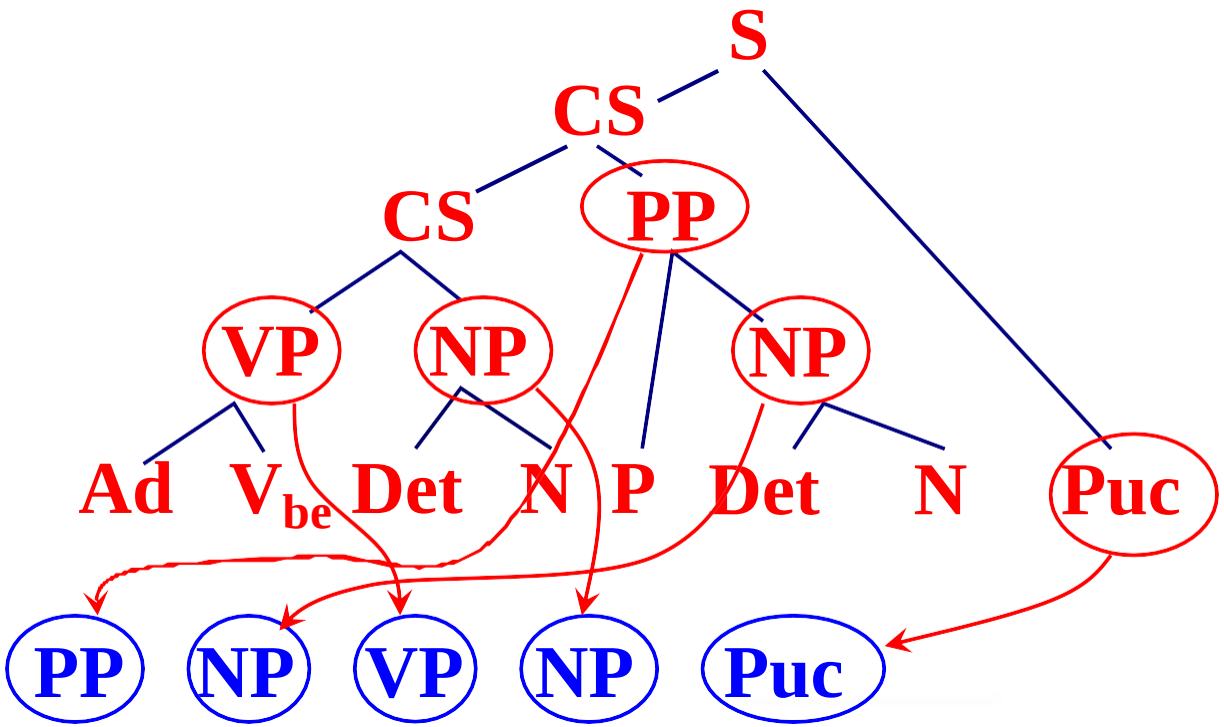
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



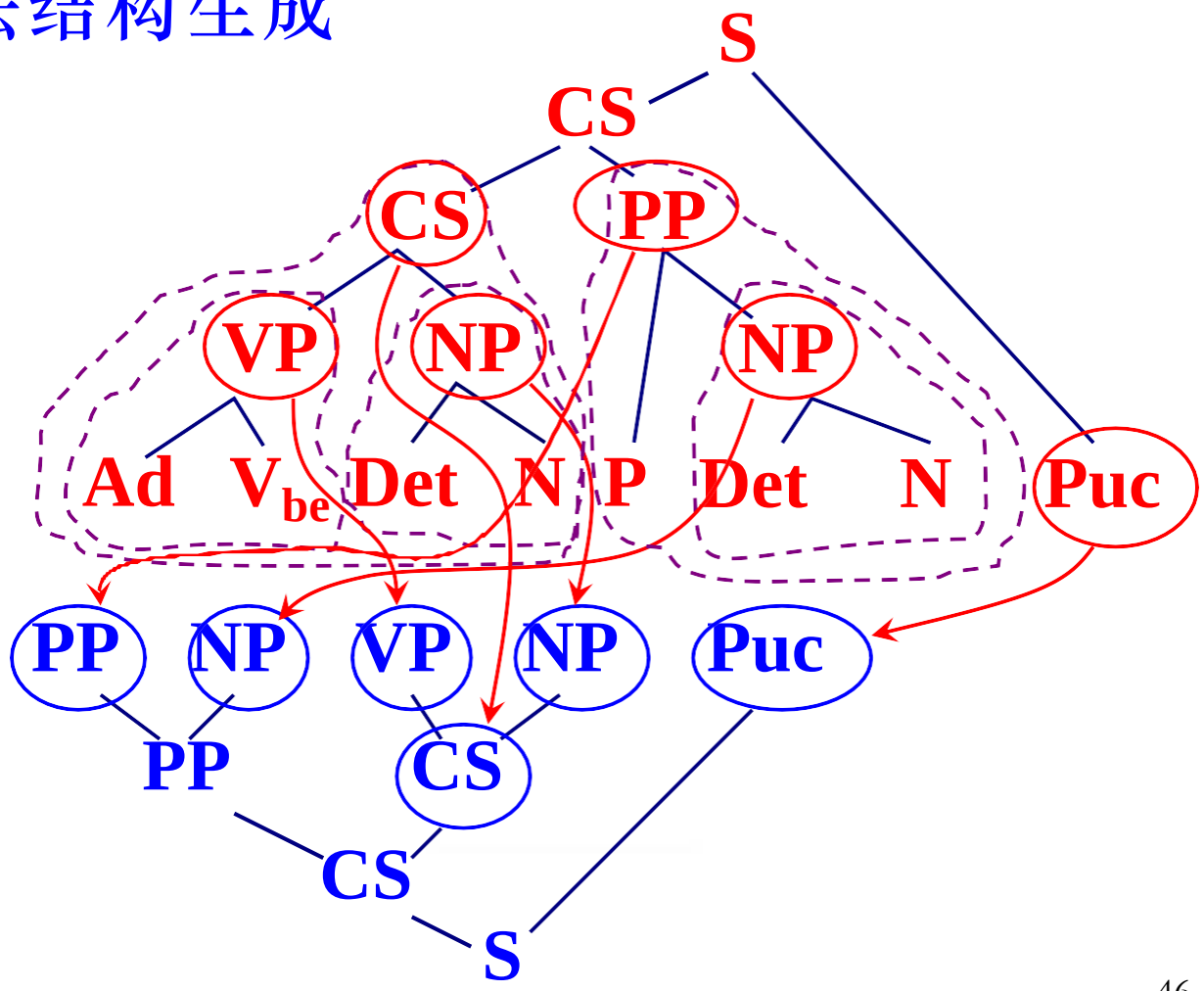
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



基本翻译方法(二)：基于规则

■Step 4:译文句法结构生成



基本翻译方法(二)：基于规则

- Step 5: 将源语言词汇翻译成目标语言词汇

# there	Ad: 在那里
# be	V_{be}: 是
# there be	VP: 在…有
# a	Det: 一，一个，一本…
# book	N: 书，书籍; V: 预订

- Step 6: 译文词法处理和目标语言句子生成：

  在桌子上有一本书。

基本翻译方法(二)：基于规则

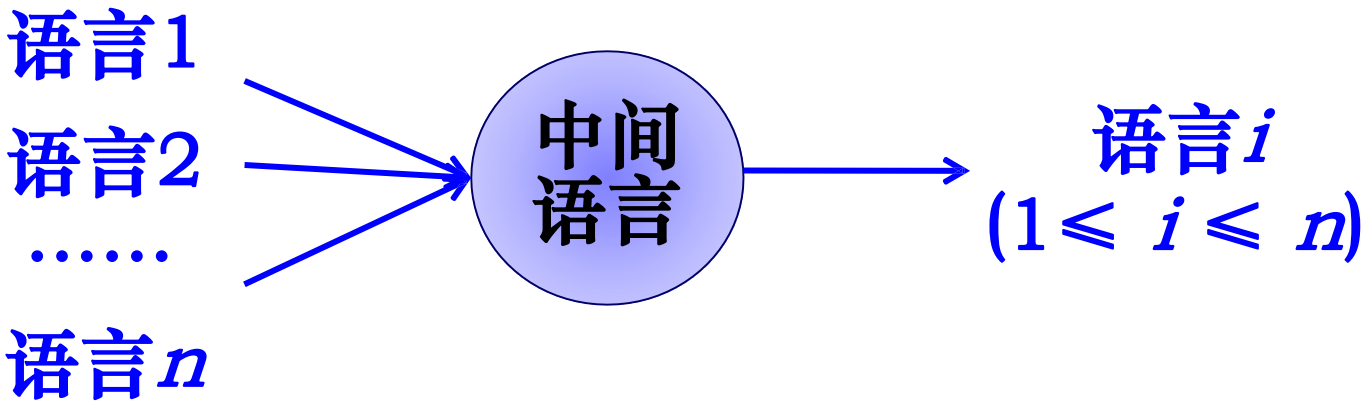
❑ 对基于规则的翻译方法的评价：

- 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.
- 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.

基本翻译方法(三)：基于中间语言

❑ 基于中间语言的翻译方法

- 方法：输入语句→中间语言→ 翻译结果



基本翻译方法(三)：基于中间语言

❑ 关于中间语言的定义

- 国际先进语音翻译研究联盟(C-STAR)定义的中间转换格式(Interchange Format)
- 日本东京联合国大学(United Nations University)提出的通用网络语言(Universal Networking Language)

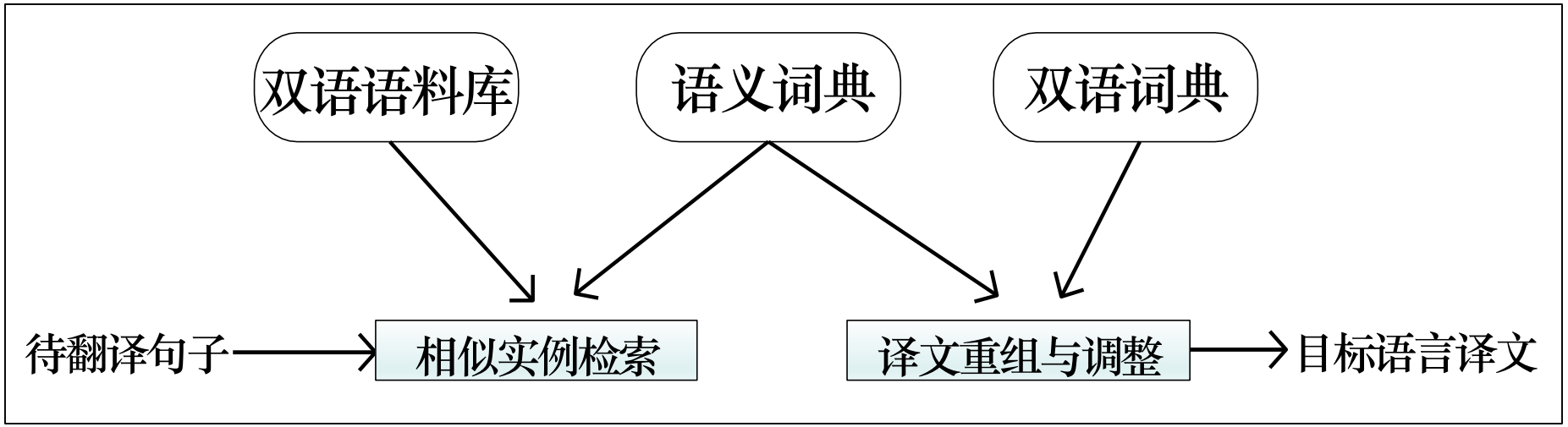
基本翻译方法(三)：基于中间语言

❑ 对基于中间语言的翻译方法评价：

- 优点：中间语言的设计可以不考虑具体的翻译语言对，因此，该方法尤其适合多语言之间的互译。
- 弱点：如何定义和设计中间语言的表达方式，以及如何维护并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面，都面临若干困难。

基本翻译方法(四)：基于事例

- 基于事例(实例)的翻译方法(Example-based)
  - 方法：输入语句→与事例相似度比较→ 翻译结果
  - 资源：大规模事例库



基本翻译方法(四)：基于事例

❑ 对基于实例的翻译方法评价：

- 优点：不要求源语言句子必须符合语法规定，翻译机制一般不需要对源语言句子做深入分析.
- 弱点：两个不同的句子之间的相似性往往难以把握； 系统往往难以处理事例库中没有记录的陌生的语言现象，而且当事例库达到一定规模时，其事例检索的效率较低.

基本翻译方法

❑ 其它翻译方法

- ❖ 统计翻译方法(statistical method)
- ❖ 基于神经网络(neural network)的翻译方法

小结

- 机器翻译的产生与发展
- 机器翻译研究现状
- 机器翻译基本方法
  - 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
  - 基于事例的翻译方法

Thank you!

权小军 中山大学数据科学与计算机学院