

自然语言处理

Natural Language Processing

权小军 教授

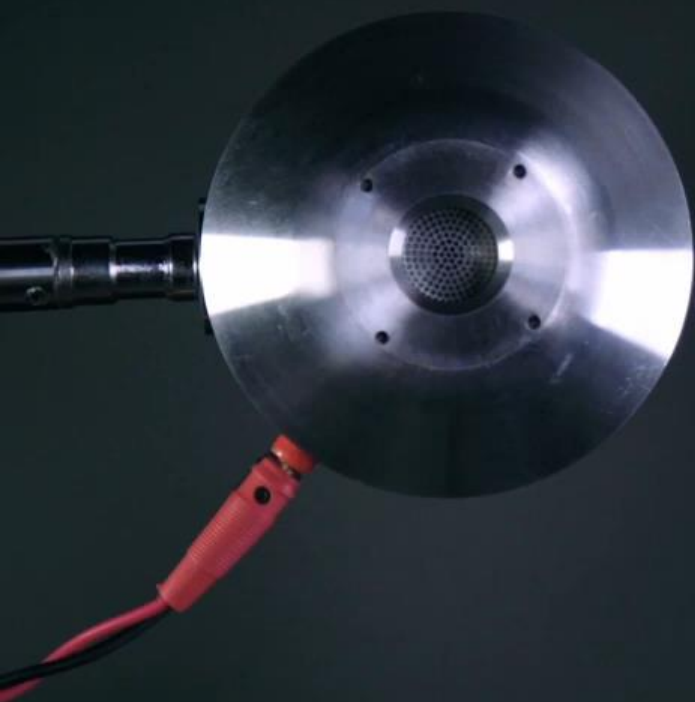
中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

展示（一）



展示 (二)



以下將播放網上不同
口音的英語作翻譯示範

展示 (三)



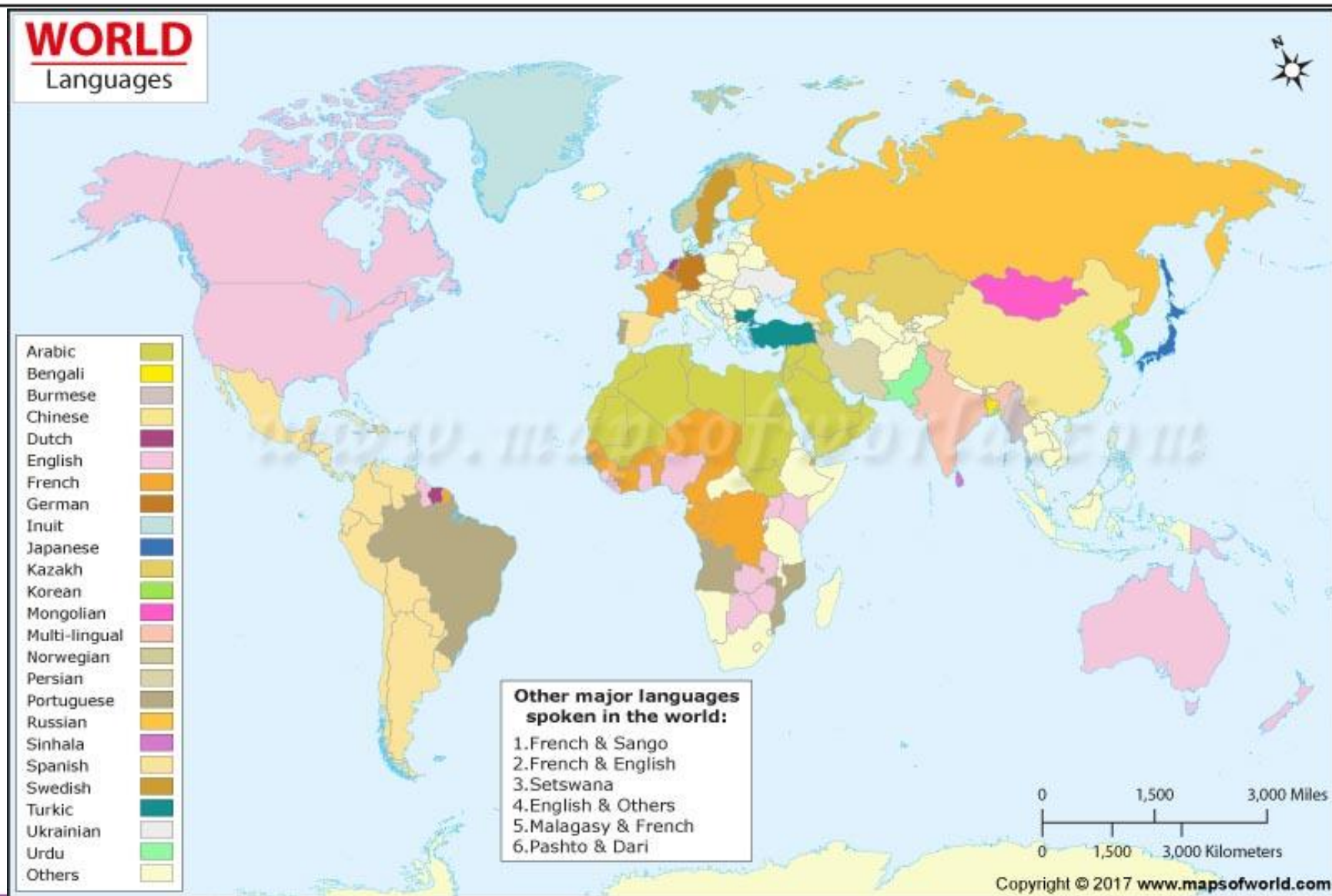
Lecture 12: 机器翻译（上）

12.1: 机器翻译概论

巴别塔 (Tower of Babel)



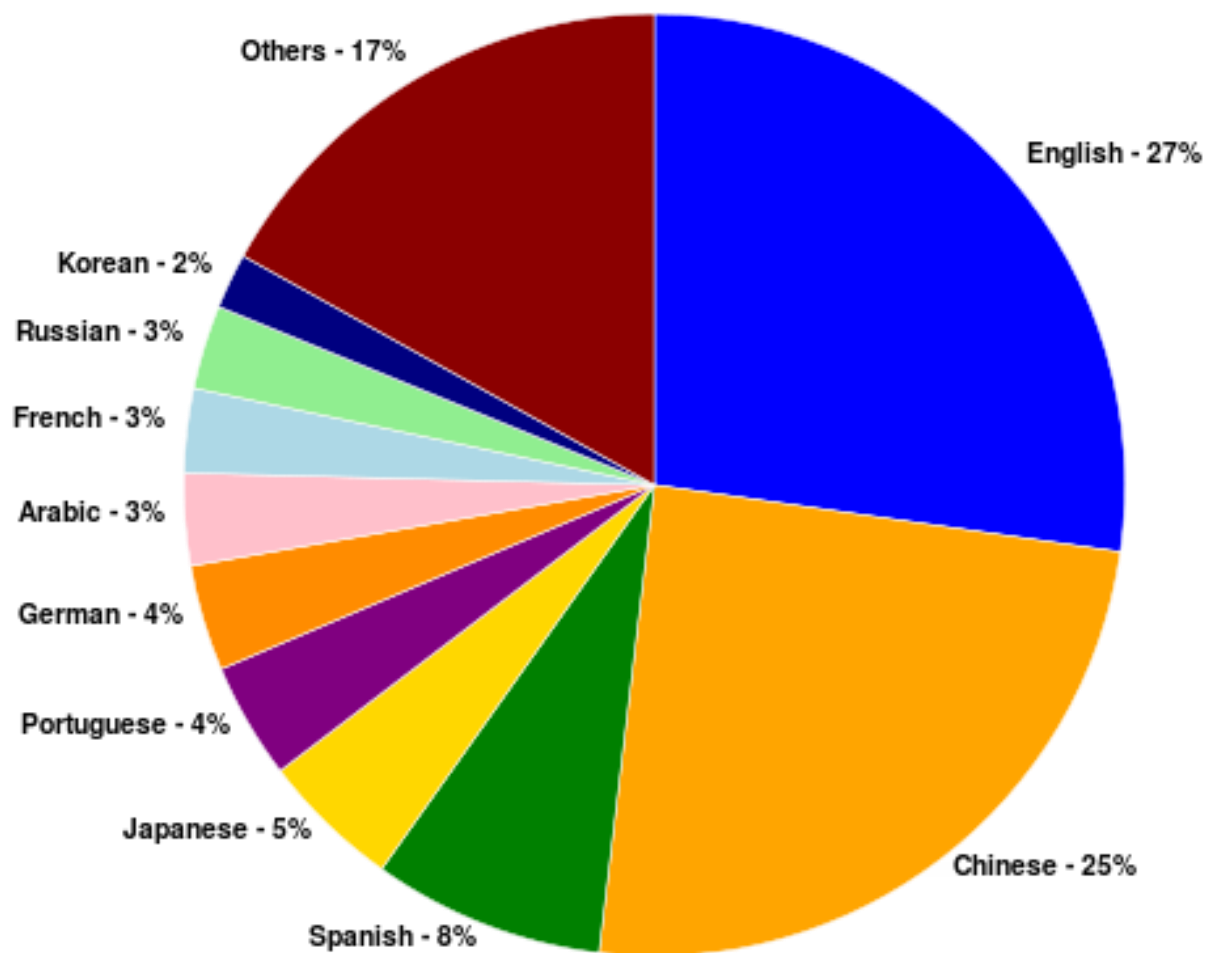
世界语言地图



中国语言地图



互联网用户



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；



引言

- ❖ 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

DETECT LANGUAGE

ENGLISH

SPANISH

FRENCH

^

↔

ENGLISH

SPANISH

ARABIC

▼

←

Search languages

✓ Detect language 🚩	Danish	Hmong	Lithuanian	Romanian	Telugu
Afrikaans	Dutch	Hungarian	Luxembourgish	Russian	Thai
Albanian	English	Icelandic	Macedonian	Samoan	Turkish
Amharic	Esperanto	Igbo	Malagasy	Scots Gaelic	Turkmen
Arabic	Estonian	Indonesian	Malay	Serbian	Ukrainian
Armenian	Filipino	Irish	Malayalam	Sesotho	Urdu
Azerbaijani	Finnish	Italian	Maltese	Shona	Uyghur
Basque	French	Japanese	Maori	Sindhi	Uzbek
Belarusian	Frisian	Javanese	Marathi	Sinhala	Vietnamese
Bengali	Galician	Kannada	Mongolian	Slovak	Welsh
Bosnian	Georgian	Kazakh	Myanmar (Burmese)	Slovenian	Xhosa
Bulgarian	German	Khmer	Nepali	Somali	Yiddish
Catalan	Greek	Kinyarwanda	Norwegian	Spanish	Yoruba
Cebuano	Gujarati	Korean	Odia (Oriya)	Sundanese	Zulu
Chichewa	Haitian Creole	Kurdish (Kurmanji)	Pashto	Swahili	
Chinese	Hausa	Kyrgyz	Persian	Swedish	
Corsican	Hawaiian	Lao	Polish	Tajik	
Croatian	Hebrew	Latin	Portuguese	Tamil	
Czech	Hindi	Latvian	Punjabi	Tatar	

108 种语言

引言



203 种语言

12.2: 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 with 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。

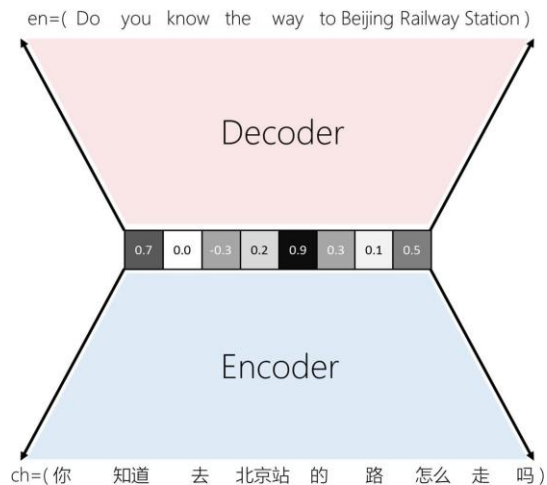
机器翻译的产生与发展

□ 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)

进一步爆发

- 时至今日，神经机器翻译已经成为新的范式，大有全面替代统计机器翻译之势



机器翻译已经成为巨头们的必争之地（图：WMT19参赛队伍）

有效应用场景

□ 专利翻译

- 对上千万专利文档进行翻译的人工代价 几乎为天文数字
- 机器翻译（集群） 仅需几天

□ 网络环境海量数据的翻译

- 情报部门对信息的截获与侦测
- 数据量超大，机器翻译几乎是唯一的选 择

□ 翻译结果后编辑

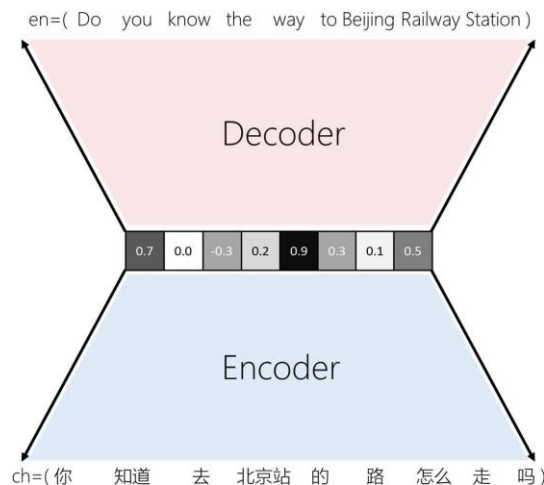
- 在机器翻译的结果上人工修改，生成可交付的翻译结果
- 市场很大

□ 全球化

- 一个公司的产品，在各个国家推广
- 文档翻译的人工成本很高，而机器翻译可以很好的完成

进一步爆发

- 时至今日，神经机器翻译已经成为新的范式，大有全面替代统计机器翻译之势



机器翻译已经成为巨头们的必争之地（图：WMT19参赛队伍）

12.3: 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 (2019)

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。

"Wang Lushan waterfall"

The incense burner in Rizhao produces purple smoke,
From a distance, you can see the waterfall hanging in
QianChuan.
Flying down three thousand feet,
It is suspected that the Milky way has set nine days.

百度翻译 (2020)

机器翻译的困难



机器翻译的困难



机器翻译的困难



机器翻译的困难



官网

抗疫行动

翻译 API

同传

上新啦

视频翻译

人工翻译

插

200+ 语种

检测到中文(简体)



英语

翻 译

人工翻译

通用领域

请勿戏水



Don't play in the water

12.4: 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 Translate
- 百度翻译

□ 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。[\(2015.4.28\)](#)

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。[\(2016.5.1\)](#)

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。[\(2017.4.16\)](#)

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

- ❖ 北京向马尼拉提出第三次严正交涉，并警告称，很难对一个岛屿的领土僵局持乐观态度。当局说，他们已经为马尼拉局势的任何升级做好了准备。[\(2020.12.01\)](#)

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.

12.5: 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. → 我(I) 喜欢(like) 玛丽(Mary).

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：

How are you ? 怎么 是 你 ?

How old are you ? 怎么 老 是 你 ?

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

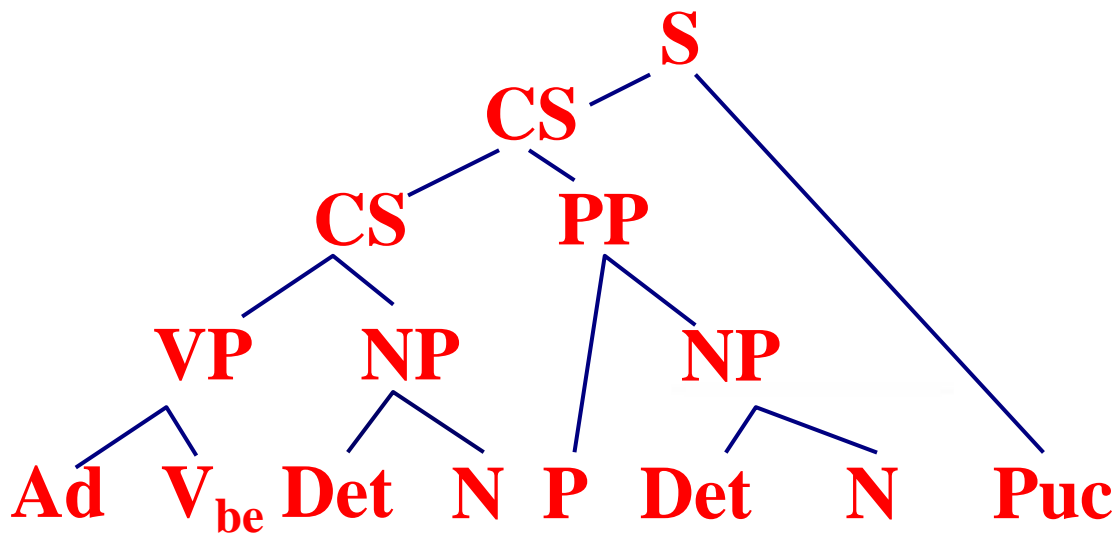
基本翻译方法(二): 基于规则

给定源语言句子: There is a book on the desk.

■ Step 1: 词法分析:

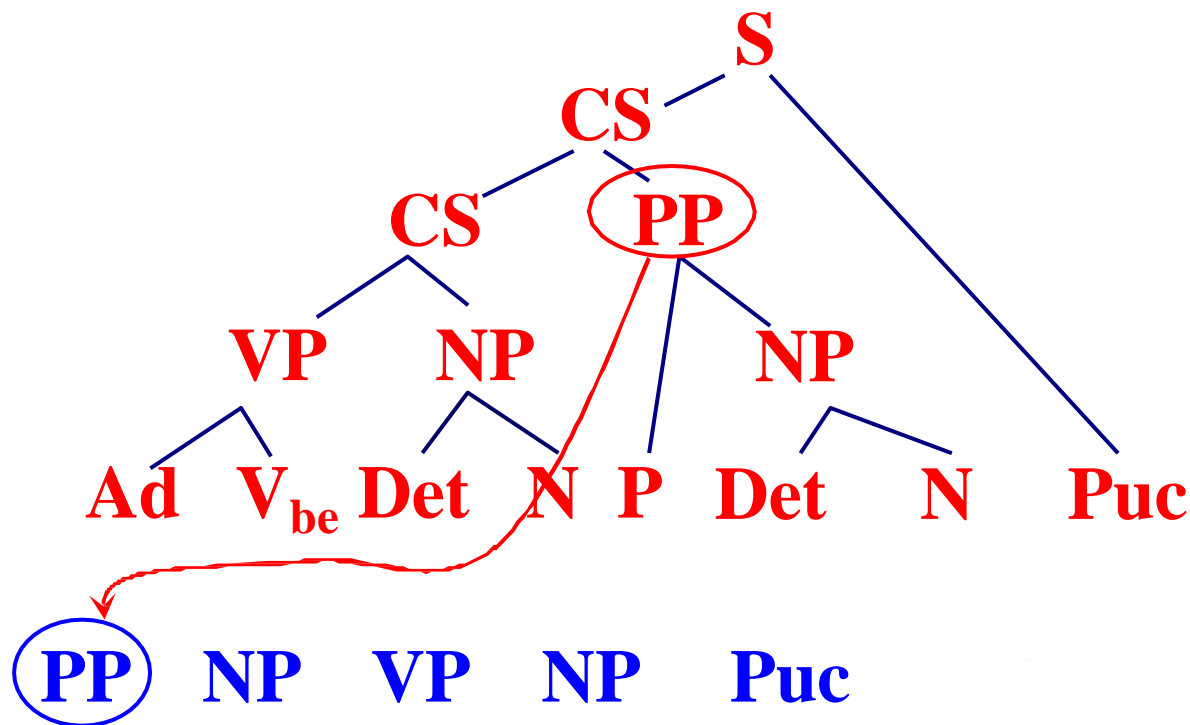
There/**Ad** is/**V_{be}** a/**Det** book/**N** on/**P** the/**Det** desk/**N** ./**Puc**

■ Step 2: 利用句法规则进行句法结构分析:



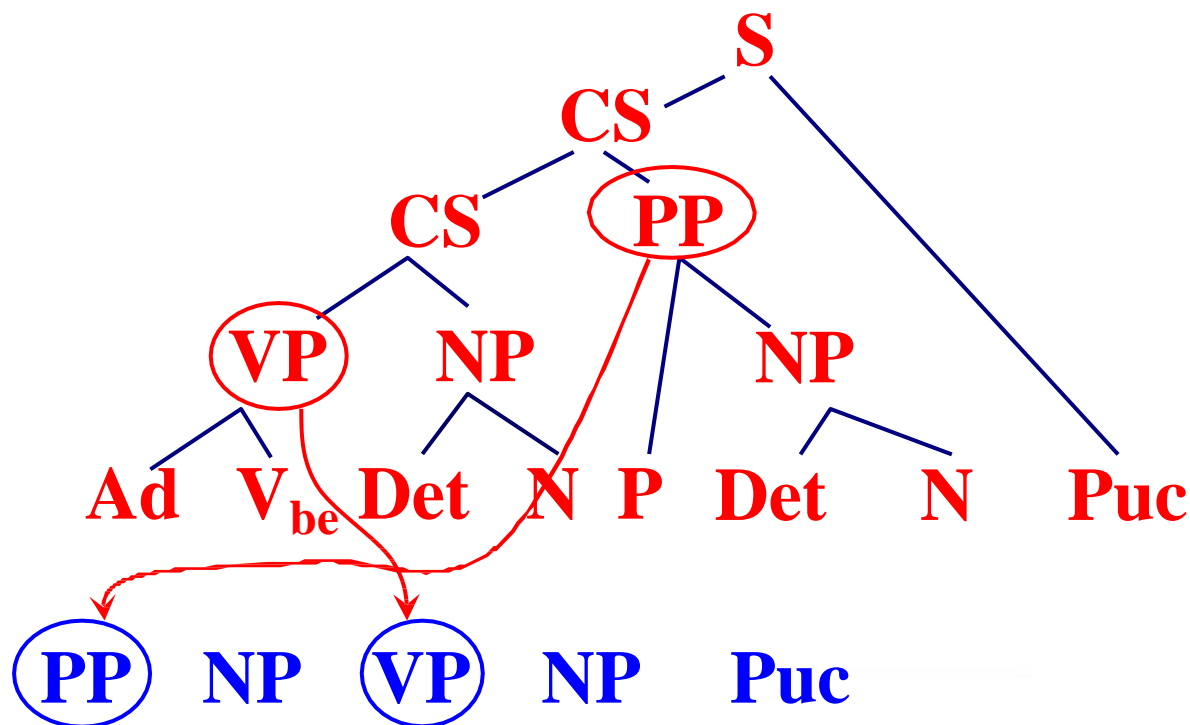
基本翻译方法(二): 基于规则

■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



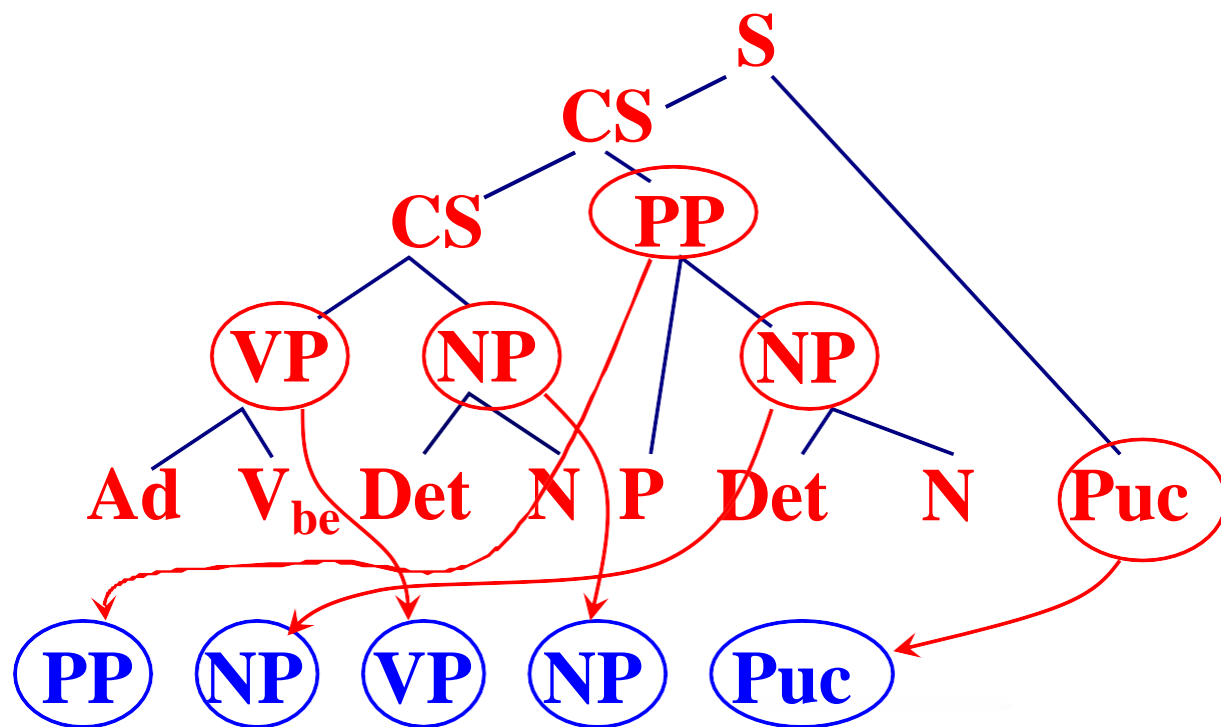
基本翻译方法(二): 基于规则

■ 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



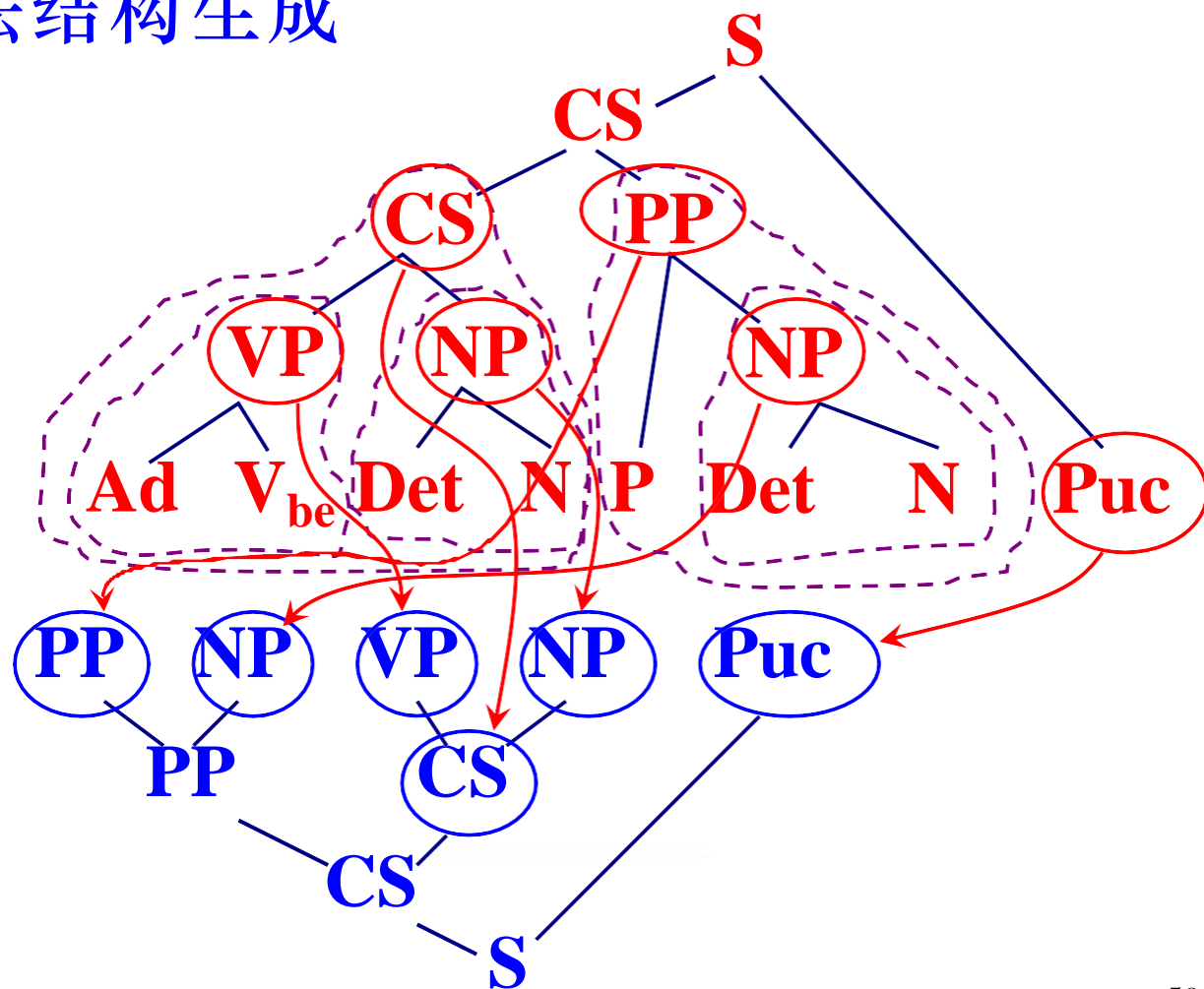
基本翻译方法(二): 基于规则

■ 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

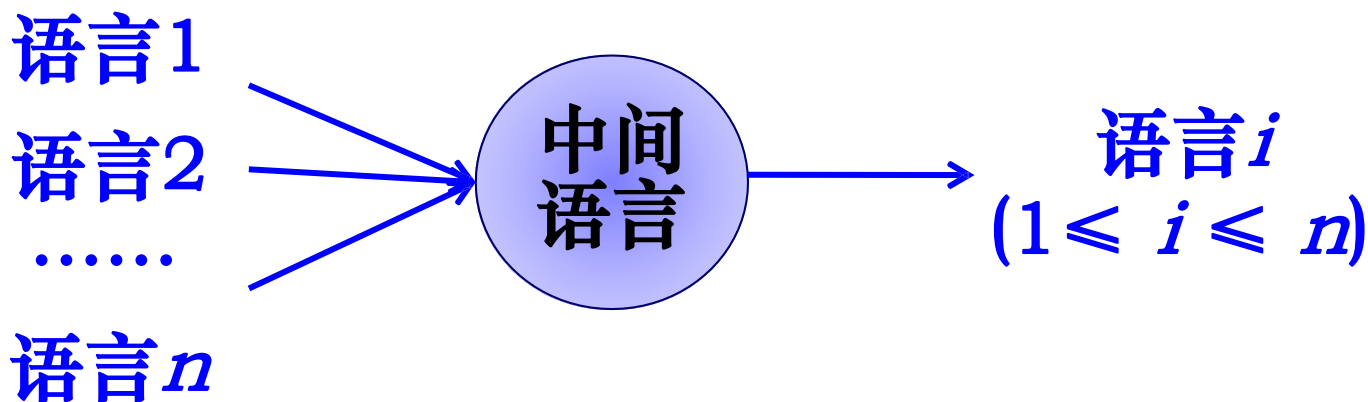
□ 对基于规则的翻译方法的评价:

- 优点: 可以较好地保持原文的结构, 产生的译文结构与原文的结构关系密切, 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 \rightarrow 中间语言 \rightarrow 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

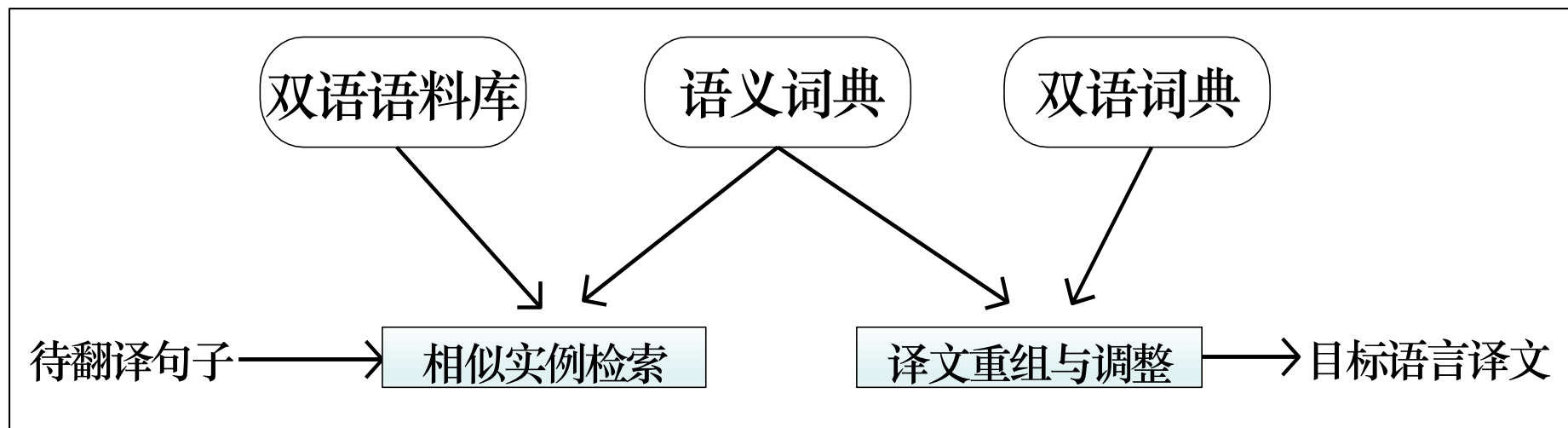
□ 对基于中间语言的翻译方法评价:

- 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): 基于实例

□ 基于实例的翻译方法(Example-based)

- 方法: 输入语句 → 与实例相似度比较 → 翻译结果
- 资源: 大规模事例库



基本翻译方法(四)：基于实例

□ 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则，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。

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么 时候 开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我 对 他 感到 高兴'
译 = 'I am happy with him'
- ...

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我 对 他 感到 高兴'
译 = 'I am happy with him'
- ...

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

我 对 你 感 到 满 意

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我 对 他 感到 高兴'
译 = 'I am happy with him'
- ...

相似度=0

我 对 你 感 到 满 意

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我 对 他 感到 高兴' <
译 = 'I am happy with him'
- ...

相似度=0.6

> 我 对 你 感 到 满 意

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我对他感到高兴'
译 = 'I am happy with him'
- ...

相似实例

我 对 你 感 到 满 意

我 对 他 感 到 高 兴

I am happy with him

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我对他感到高兴'
译 = 'I am happy with him'
- ...

我 对 你 感 到 满 意

↑
不匹配
↓

我 对 他 感 到 高 兴

相似实例

I am happy with him

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始?'
译 = 'When will it start?'
- 2: 源 = '我对他感到高兴'
译 = 'I am happy with him'
- ...

我 对 你 感 到 满 意



不匹配



不匹配

我 对 他 感 到 高 兴

相似实例

I am happy with him

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我对他感到高兴'
译 = 'I am happy with him'
- ...

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

我 对 你 感 到 满 意



不匹配



不匹配

我 对 他 感 到 高 兴

相似实例

I am happy with him



用 '你' 替换 '他'

我 对 你 感 到 高 兴

I am happy with you

基于实例的机器翻译

资源1：翻译实例库

- 1: 源 = '什么时候开始 ?'
译 = 'When will it start ?'
- 2: 源 = '我对他感到高兴'
译 = 'I am happy with him'
- ...

相似实例

资源2：翻译词典

- 1: 我 → I | me
- 2: 你 → you
- 3: 满意 → satisfy | satisfied ...
- ...

我 对 你 感 到 满 意

不匹配

不匹配

我 对 他 感 到 高 兴

I am happy with him

用 '你' 替换 '他'

我 对 你 感 到 高 兴

I am happy with you

用 '满意' 替换 '高兴'

我 对 你 感 到 满 意

I am **satisfied** with you

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

基于统计的方法

统计机器翻译：利用统计模型从双/单语语料中自动学习翻译知识，之后这些知识对新句子进行翻译

❖ 建模 + 学习知识 + 利用这些知识进行翻译

基于统计的方法

资源1:双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

基于统计的方法

资源1:双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\text{Pr}(\text{我} \rightarrow \text{I}) = 0.7$
 $\text{Pr}(\text{我} \rightarrow \text{me}) = 0.3$
 $\text{Pr}(\text{你} \rightarrow \text{you}) = 0.9$
 $\text{Pr}(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\text{Pr}(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\text{Pr}(\text{我} \rightarrow \text{I}) = 0.7$
 $\text{Pr}(\text{我} \rightarrow \text{me}) = 0.3$
 $\text{Pr}(\text{你} \rightarrow \text{you}) = 0.9$
 $\text{Pr}(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\text{Pr}(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\text{Pr}(\text{我} \rightarrow \text{I}) = 0.7$
 $\text{Pr}(\text{我} \rightarrow \text{me}) = 0.3$
 $\text{Pr}(\text{你} \rightarrow \text{you}) = 0.9$
 $\text{Pr}(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\text{Pr}(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\text{Pr}(\text{I}) = 0.0001$
 $\text{Pr}(\text{I} \rightarrow \text{am}) = 0.623$
 $\text{Pr}(\text{I} \rightarrow \text{was}) = 0.21$
...

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\text{Pr}(\text{我} \rightarrow \text{I}) = 0.7$
 $\text{Pr}(\text{我} \rightarrow \text{me}) = 0.3$
 $\text{Pr}(\text{你} \rightarrow \text{you}) = 0.9$
 $\text{Pr}(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\text{Pr}(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\text{Pr}(\text{I}) = 0.0001$
 $\text{Pr}(\text{I} \rightarrow \text{am}) = 0.623$
 $\text{Pr}(\text{I} \rightarrow \text{was}) = 0.21$
...

翻译引擎

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\text{Pr}(\text{我} \rightarrow \text{I}) = 0.7$
 $\text{Pr}(\text{我} \rightarrow \text{me}) = 0.3$
 $\text{Pr}(\text{你} \rightarrow \text{you}) = 0.9$
 $\text{Pr}(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\text{Pr}(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

我 对 你 感 到 满 意

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\text{Pr}(\text{I}) = 0.0001$
 $\text{Pr}(\text{I} \rightarrow \text{am}) = 0.623$
 $\text{Pr}(\text{I} \rightarrow \text{was}) = 0.21$
...

翻译引擎

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\Pr(\text{我} \rightarrow \text{I}) = 0.7$
 $\Pr(\text{我} \rightarrow \text{me}) = 0.3$
 $\Pr(\text{你} \rightarrow \text{you}) = 0.9$
 $\Pr(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\Pr(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\Pr(\text{I}) = 0.0001$
 $\Pr(\text{I} \rightarrow \text{am}) = 0.623$
 $\Pr(\text{I} \rightarrow \text{was}) = 0.21$
...

我 对 你 感 到 满 意

翻译假设

I to you happy
You satisfied
I satisfied with you
I'm satisfied with you
I satisfied you, what
You can have it
You and me



枚举所有可能

翻译引擎

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\Pr(\text{我} \rightarrow \text{I}) = 0.7$
 $\Pr(\text{我} \rightarrow \text{me}) = 0.3$
 $\Pr(\text{你} \rightarrow \text{you}) = 0.9$
 $\Pr(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\Pr(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\Pr(\text{I}) = 0.0001$
 $\Pr(\text{I} \rightarrow \text{am}) = 0.623$
 $\Pr(\text{I} \rightarrow \text{was}) = 0.21$
...

我 对 你 感 到 满 意

翻译假设	概率
I to you happy	0.01
You satisfied	0.02
I satisfied with you	0.10
I'm satisfied with you	0.46
I satisfied you, what	0.23
You can have it	0.01
You and me	0.02

翻译引擎

计算翻译可能性

基于统计的方法

资源1: 双语平行语料

- 1: 源='他 在 哪 ?'
译='Where is he ?'
- 2: 源='我 真 高兴'
译='I'm so happy'
- 3: 源='出发 !'
译='Let's go!'
- ...

学习

翻译模型

$\Pr(\text{我} \rightarrow \text{I}) = 0.7$
 $\Pr(\text{我} \rightarrow \text{me}) = 0.3$
 $\Pr(\text{你} \rightarrow \text{you}) = 0.9$
 $\Pr(\text{开心} \rightarrow \text{happy}) = 0.5$
 $\Pr(\text{满意} \rightarrow \text{satisfied}) = 0.4$
...

资源2: 单语语料

- 1: What is your dream?
- 2: Are you fulfilled ?
- 3: Yes, you are right .
- ...

学习

语言模型

$\Pr(\text{I}) = 0.0001$
 $\Pr(\text{I} \rightarrow \text{am}) = 0.623$
 $\Pr(\text{I} \rightarrow \text{was}) = 0.21$
...

我 对 你 感 到 满 意

翻译假设	概率
I to you happy	0.01
You satisfied	0.02
I satisfied with you	0.10
I'm satisfied with you	0.46
I satisfied you, what	0.23
You can have it	0.01
You and me	0.02

翻译引擎

计算翻译可能性

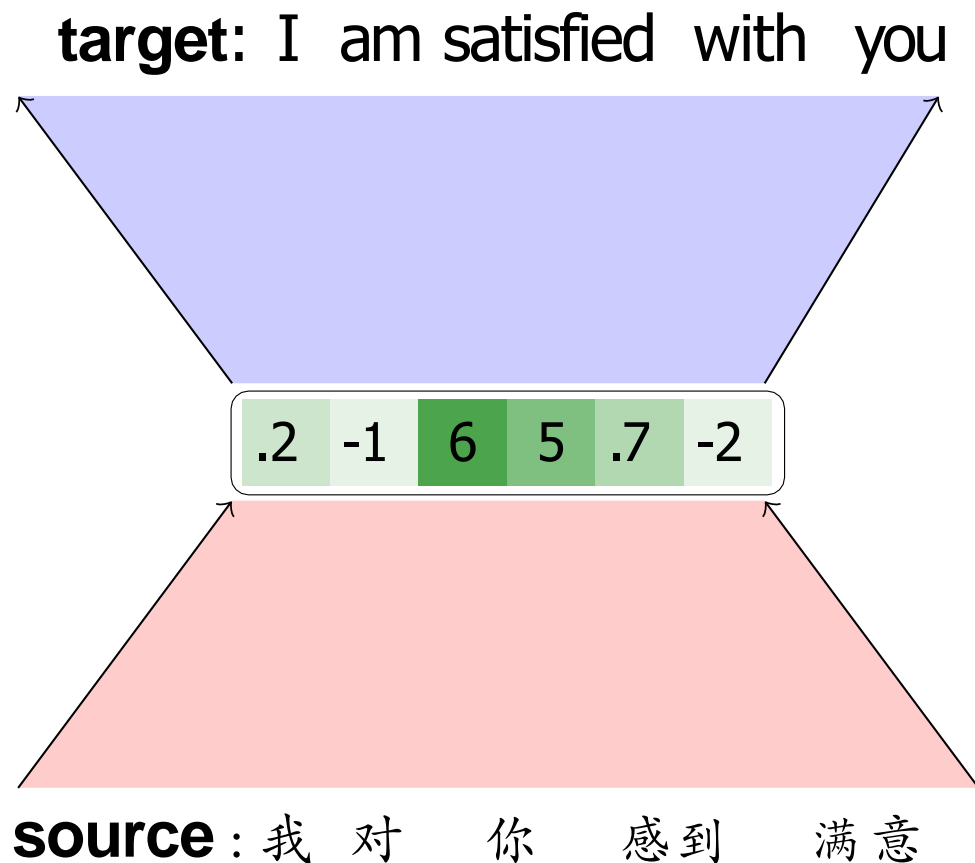
基于神经网络的方法

神经机器翻译：把字符串表示成实数向量，翻译被看做序列到序列转化。这个过程由**编码器-解码器**网络计算，不需要依赖人工先验或者语言学规则。

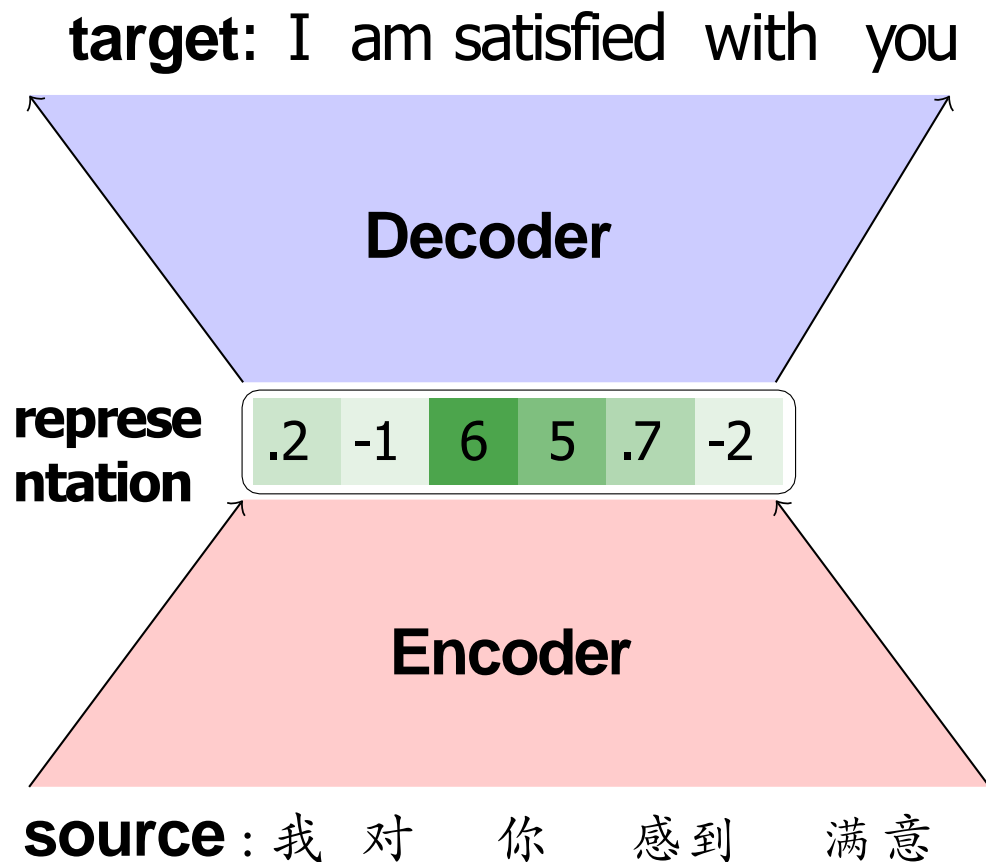
基于神经网络的方法



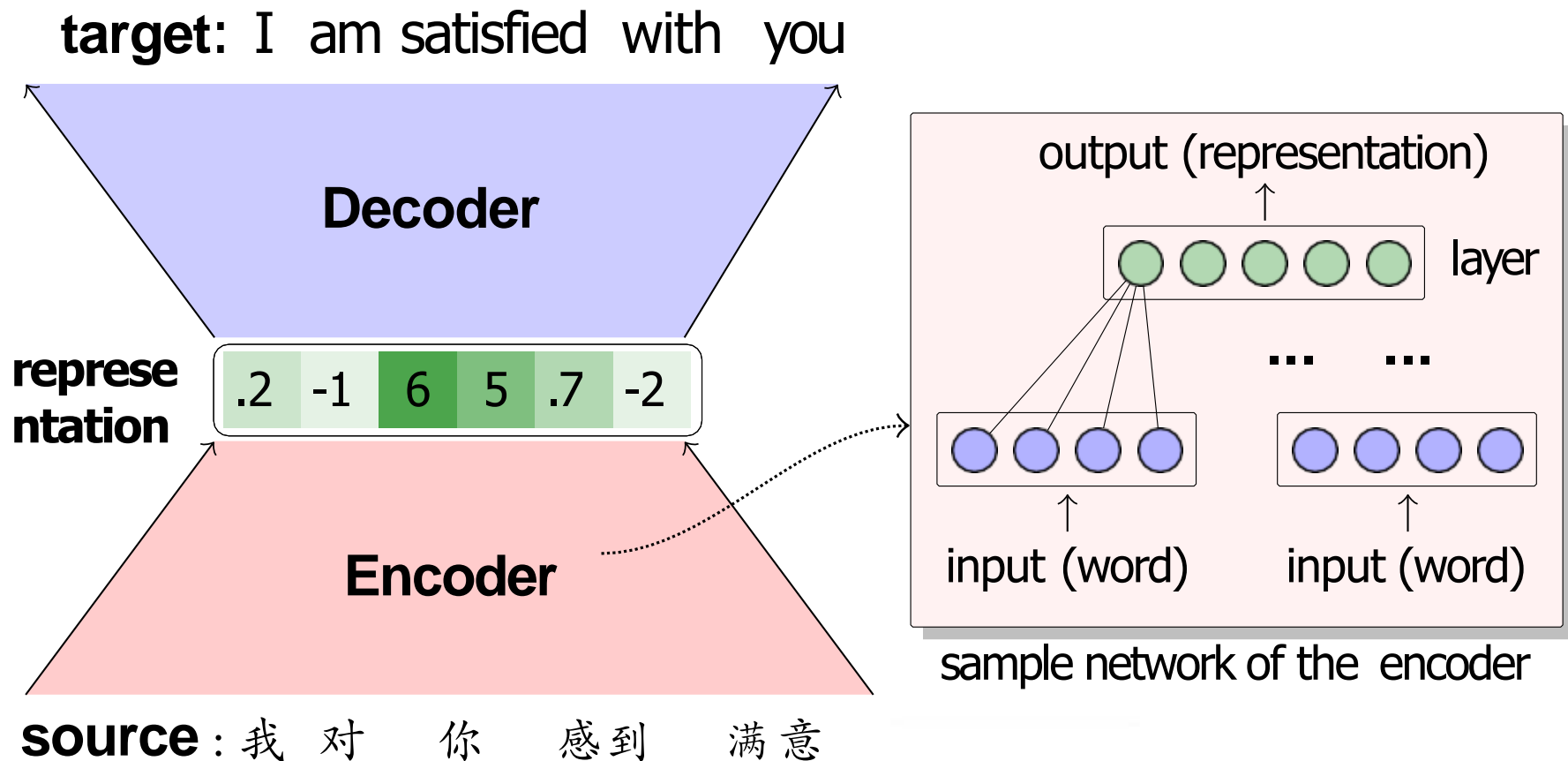
基于神经网络的方法



基于神经网络的方法



基于神经网络的方法



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