机器翻译简介

背景 + 概念 + 方法概述

肖桐　朱靖波

xiaotong@mail.neu.edu.cn
zhujingbo@mail.neu.edu.cn

东北大学 自然语言处理实验室
http://www.nlplab.com

翻译

广义上来说，"翻译"是指把一个事物转化为另一个事物的过程。这个概念多使用在对可序列化的符号串的转化上，比如

- **程序编译**：
  从一种编程语言的代码转化到另外一种语言的代码
- **文字翻译**：
  从一种语言文字转化到另一种语言文字
- **蛋白质生物合成第一步**：
  从一种RNA分子序列转化到特定氨基酸序列

翻译

广义上来说，"翻译"是指把一个事物转化为另一个事物的过程。这个概念多使用在对可序列化的符号串的转化上，比如

- **程序编译**：
  从一种编程语言的代码转化到另外一种语言的代码
- **文字翻译**：
  从一种语言文字转化到另一种语言文字
- **蛋白质生物合成第一步**：
  从一种RNA分子序列转化到特定氨基酸序列

这里我们只关注从一种语言的文字序列翻译到另外一种语言的文字序列的过程，即自然语言的翻译

中文　　英文

我们在翻译一段文字 ⟶ We are translating a piece of text

啥是机器翻译

传统观念中，翻译是由人工完成

- 精度高
- 但是，费时费力

啥是机器翻译

传统观念中，翻译是由人工完成

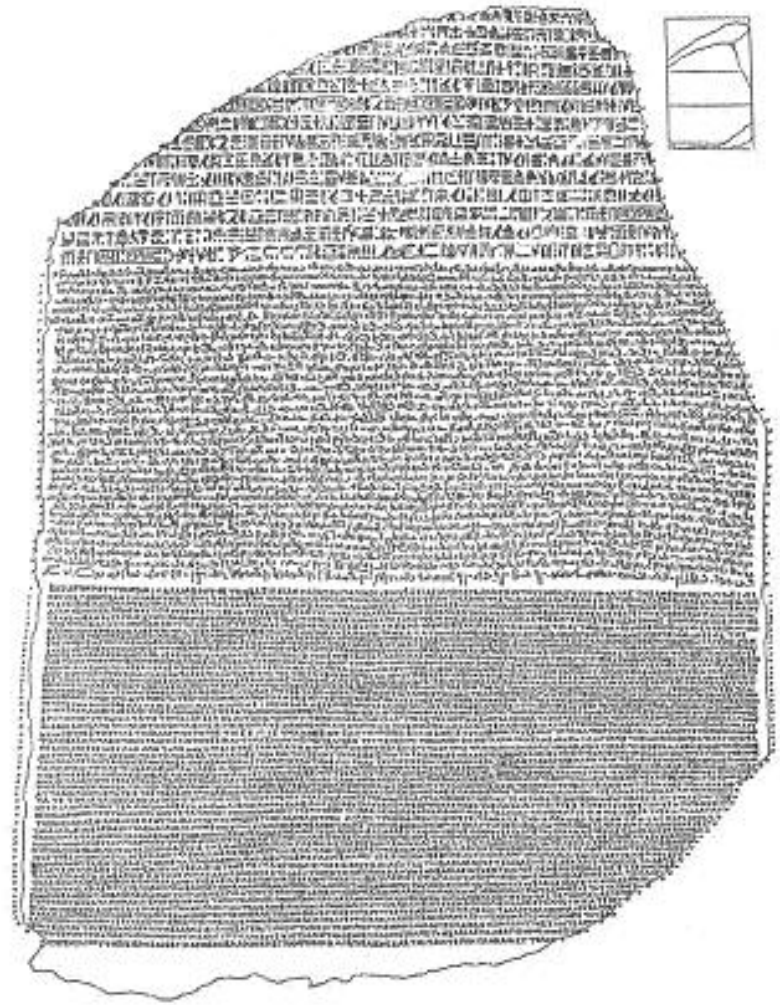
- 精度高
- 但是，费时费力

机器翻译是指由计算机完成自动翻译，而非人工

- 精度有限，有时非常差
- 但是，速度快成本极低

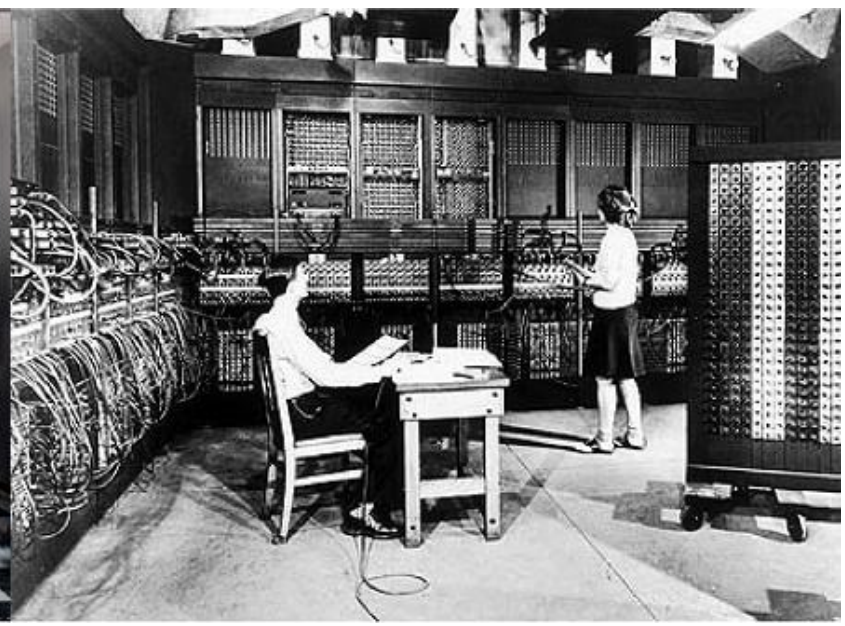
早期的(人工)翻译

- 人类文明最早的翻译记录
  - 罗塞塔石碑(Rosetta Stone)
  - 刻有埃及国王诏书
  - 石碑上用希腊文字、古埃及文字和当时的通俗体文字刻了同样的内容
- 宗教文献翻译
  - 中国的邻国在很长时间内都没有自己的文字，直到佛教传入
  - 在西方，一项最早被记录的翻译活动是将旧约圣经译成希腊语
  - 罗马圣哲杰罗姆将圣经翻译成拉丁语



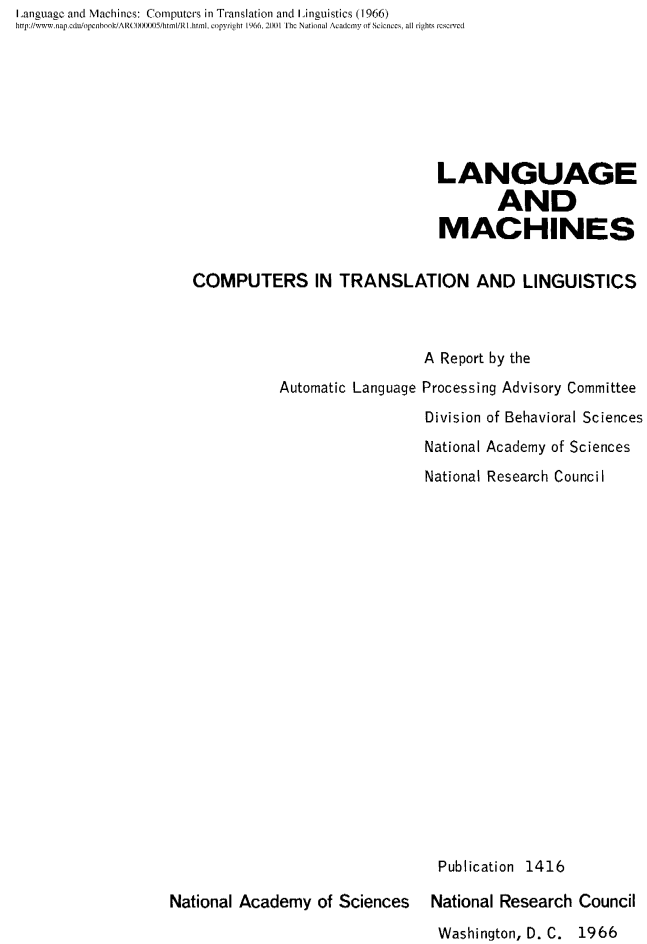
机器翻译的开始

- 20世纪30年代初，法国科学家G.B.阿尔楚尼提出了用机器来进行翻译的想法
- 40年代，美国科学家W. Weaver 和英国工程师A. D. Booth提出了利用计算机进行语言自动翻译的想法
- 1949年，W. Weaver 发表《翻译备忘录》，正式提出机器翻译的思想



受挫期

- 从20世纪50、60年代，美国和苏联两个超级大国由于政治和经济的需要也对机器翻译研究给予了相当大的重视
- 不幸的是
  - 1964年，美国科学院成立了语言自动处理咨询委员会(ALPAC委员会)，开始了为期两年的综合调查分析和测试
  - 1966年11月，该委员会公布了一个题为《语言与机器》的报告(简称ALPAC报告)
  - 该报告全面否定了机器翻译的可行性，并建议停止对机器翻译项目的资金支持



LANGUAGE AND MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

Publication 1416

National Academy of Sciences National Research Council
Washington, D. C. 1966

快速成长期

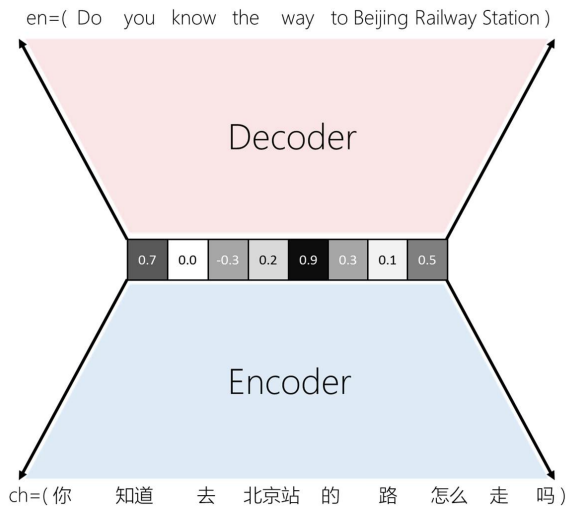
- 进入70年代后，国与国之间的语言障碍显得更为严重，传统的人工作业方式已经远远不能满足需求
- 特别是90年代后期，随着 Internet 的普遍应用，机器翻译迎来了一个新的发展机遇
  - 有关机器翻译研究的会议频繁召开
  - 各种机器翻译公司的成立
- 互联网巨头的免费自动翻译服务更是把机器翻译带进新时代

进一步爆发

- 2013年深度学习被应用于机器翻译，受到广泛关注
  - 端到端学习不依赖过多的先验假设
  - 神经网络的连续空间模型有更强的表示能力
  - 深度网络学习算法的发展和GPU等并行计算模式为训练大规模神经网络提供了可能

进一步爆发

- 2013年深度学习被应用于机器翻译，受到广泛关注
  - 端到端学习不依赖过多的先验假设
  - 神经网络的连续空间模型有更强的表示能力
  - 深度网络学习算法的发展和GPU等并行计算模式为训练大规模神经网络提供了可能
- 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）

翻译质量

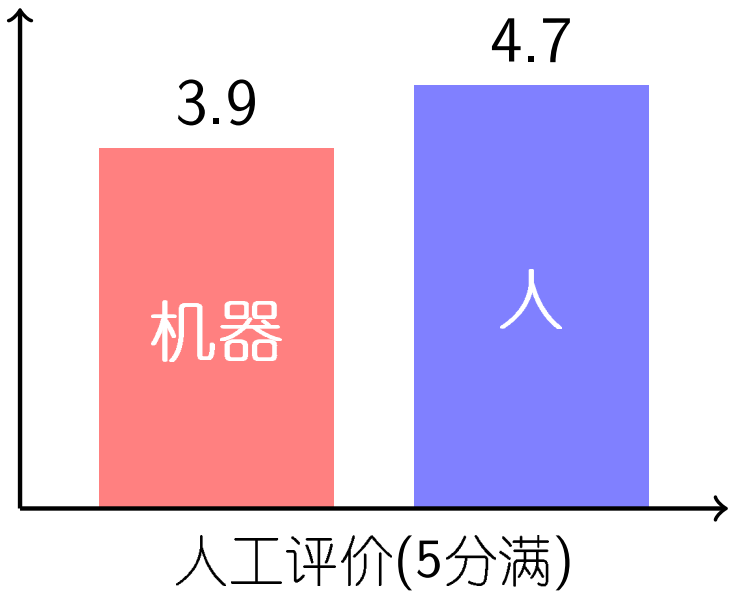
- 到了今天，机器翻译的质量究竟如何？客气地说
  - 受限条件下，机器翻译的结果可以接近人工翻译的结果
  - 开放式翻译任务，机器翻译的结果并不理想

翻译质量

- 到了今天，机器翻译的质量究竟如何？客气地说
  - 受限条件下，机器翻译的结果可以接近人工翻译的结果
  - 开放式翻译任务，机器翻译的结果并不理想
- 不客气地说，机器翻译的质量远没有达到人们的期望
  - 高精度同声传译 - 还需要打磨
  - 小说的翻译 - 可以想一想
  - 古代诗词 - ？？？不要开玩笑

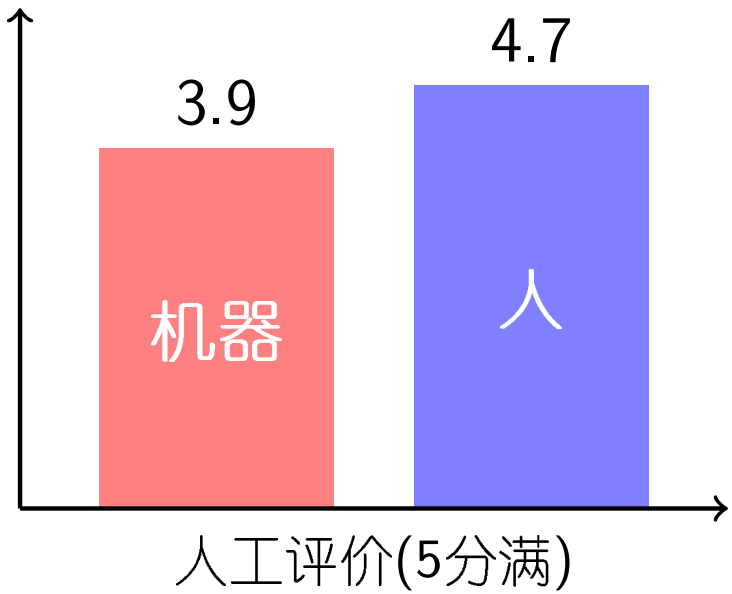
翻译质量

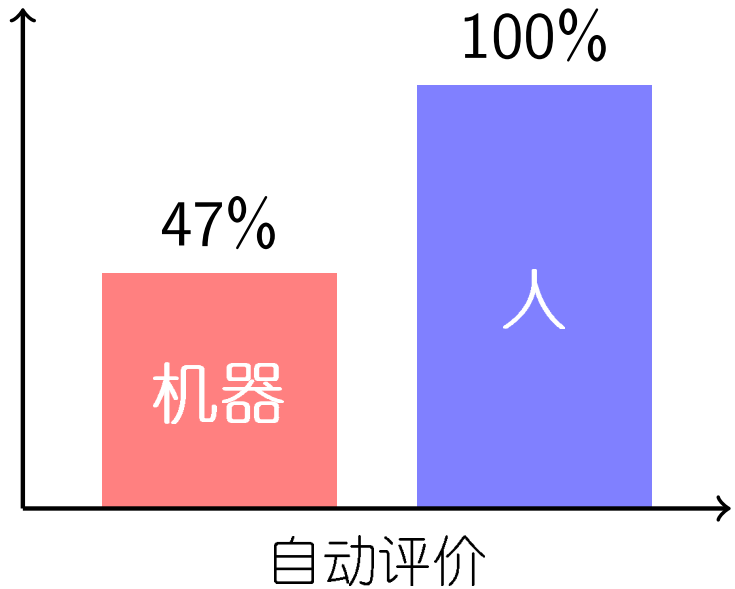
- 到了今天，机器翻译的质量究竟如何？客气地说
  - 受限条件下，机器翻译的结果可以接近人工翻译的结果
  - 开放式翻译任务，机器翻译的结果并不理想
- 不客气地说，机器翻译的质量远没有达到人们的期望
  - 高精度同声传译 - 还需要打磨
  - 小说的翻译 - 可以想一想
  - 古代诗词 - ？？？不要开玩笑
- 真实的结果 - 汉英新闻领域翻译



翻译质量

- 到了今天，机器翻译的质量究竟如何？客气地说
  - 受限条件下，机器翻译的结果可以接近人工翻译的结果
  - 开放式翻译任务，机器翻译的结果并不理想
- 不客气地说，机器翻译的质量远没有达到人们的期望
  - 高精度同声传译 - 还需要打磨
  - 小说的翻译 - 可以想一想
  - 古代诗词 - ？？？不要开玩笑
- 真实的结果 - 汉英新闻领域翻译





机器翻译没有用？

- 机器翻译一点儿用都没有？

机器翻译没有用？

- 机器翻译一点儿用都没有？

源文(约500字)

过去，曾经有一个小岛，上面住着喜悦，悲伤，知识，爱，等各种情绪。 一天，这些情绪发现这个小岛即将沉没，所以大家都陆续准备船只，打算离开小岛。 只有爱留了下来。她要坚持到最后一刻。过了几天，小岛真的要沉没了，爱想找别人帮助。 此时，财富正好乘着一艘大船经过。 爱说：“财富，你能不能带我离开？”...

机器翻译

In the past, there has been a small island, on top of joy, sadness, knowledge, love, all kinds of emotions. One day, these emotions that this is a small island, so we all have to prepare for the vessel to leave the island. The only love to stay. She wanted to go to the last minute. After a few days, the small island is to be sunk, would like to seek help from the others. At that time, the wealth by a boat. Wealth, said: "You can take me to ...

机器翻译没有用？

- 机器翻译一点儿用都没有？

人工翻译

Once upon a time there was a small island where lived all kinds of emotions like JOY, SADNESS, KNOWLEDGE and LOVE. One day, these emotions found that the island was sinking, so one by one they prepared the boat and planned to leave. None but LOVE chose to stay there. She was determined to persist till the last moment. A few days later, almost the whole island sunk into the sea, and LOVE had to seek for help. At that moment, WEALTH was ...

机器翻译

In the past, there has been a small island, on top of joy, sadness, knowledge, love, all kinds of emotions. One day, these emotions that this is a small island, so we all have to prepare for the vessel to leave the island. The only love to stay. She wanted to go to the last minute. After a few days, the small island is to be sunk, would like to seek help from the others. At that time, the wealth by a boat. Wealth, said: "You can take me to ...

机器翻译没有用？

- 机器翻译一点儿用都没有？
  - 确实，二者有差距

人工翻译

Once upon a time there was a small island where lived all kinds of emotions like JOY, SADNESS, KNOWLEDGE and LOVE. One day, these emotions found that the island was sinking, so one by one they prepared the boat and planned to leave. None but LOVE chose to stay there. She was determined to persist till the last moment. A few days later, almost the whole island sunk into the sea, and LOVE had to seek for help. At that moment, WEALTH was ...

机器翻译

In the past, there has been a small island, on top of joy, sadness, knowledge, love, all kinds of emotions. One day, these emotions that this is a small island, so we all have to prepare for the vessel to leave the island. The only love to stay. She wanted to go to the last minute. After a few days, the small island is to be sunk, would like to seek help from the others. At that time, the wealth by a boat. Wealth, said: "You can take me to ...

机器翻译没有用？

- 机器翻译一点儿用都没有？
  - 确实，二者有差距
  - 如果考虑速度 - 哦？

人工翻译 - 耗时20分钟

Once upon a time there was a small island where lived all kinds of emotions like JOY, SADNESS, KNOWLEDGE and LOVE. One day, these emotions found that the island was sinking, so one by one they prepared the boat and planned to leave. None but LOVE chose to stay there. She was determined to persist till the last moment. A few days later, almost the whole island sunk into the sea, and LOVE had to seek for help. At that moment, WEALTH was ...

机器翻译 - 耗时2秒钟

In the past, there has been a small island, on top of joy, sadness, knowledge, love, all kinds of emotions. One day, these emotions that this is a small island, so we all have to prepare for the vessel to leave the island. The only love to stay. She wanted to go to the last minute. After a few days, the small island is to be sunk, would like to seek help from the others. At that time, the wealth by a boat. Wealth, said: "You can take me to ...

机器翻译没有用？

- 机器翻译一点儿用都没有？
  - 确实，二者有差距
  - 如果考虑速度 - 哦？
  - 换个情况：如果我们要翻译100万篇这样的文档
    - 人工的成本无法想象，消耗的时间更是难以计算
    - 而计算机集群只需要1天，而且只有电力消耗

人工翻译 - 耗时20分钟

Once upon a time there was a small island where lived all kinds of emotions like JOY, SADNESS, KNOWLEDGE and LOVE. One day, these emotions found that the island was sinking, so one by one they prepared the boat and planned to leave. None but LOVE chose to stay there. She was determined to persist till the last moment. A few days later, almost the whole island sunk into the sea, and LOVE had to seek for help. At that moment, WEALTH was ...

机器翻译 - 耗时2秒钟

In the past, there has been a small island, on top of joy, sadness, knowledge, love, all kinds of emotions. One day, these emotions that this is a small island, so we all have to prepare for the vessel to leave the island. The only love to stay. She wanted to go to the last minute. After a few days, the small island is to be sunk, would like to seek help from the others. At that time, the wealth by a boat. Wealth, said: "You can take me to ...

真的很有用！

1. 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

真的很有用！

1. 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
2. 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

真的很有用！

1. 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
2. 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
3. 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

真的很有用！

1. 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

2. 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

3. 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

4. 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这个工作

是否有兴趣？

- 如果你是...，那么你一定会对机器翻译感兴趣
  1. 计算机、语言学、应用数学及相关专业的科研技术人员
  2. 翻译人员且能接受利用计算机提高工作效率
  3. 所有有翻译需求的企业技术人员
  4. ...

是否有兴趣？

- 如果你是...，那么你一定会对机器翻译感兴趣
  1. 计算机、语言学、应用数学及相关专业的科研技术人员
  2. 翻译人员且能接受利用计算机提高工作效率
  3. 所有有翻译需求的企业技术人员
  4. ...
- 这里重点介绍当今主流的机器翻译原理、方法及实现技术（统计/神经机器翻译）
  1. **第一部分：基本概念 (Sections 1-2)**
     不涉及深入的机器翻译技术 - 面向所有人
  2. **第二部分：统计机器翻译 (Sections 3-4)**
     统计机器翻译的建模，进一步介绍基于词、短语、句法的模型及实现方法
  3. **第三部分：神经机器翻译 (Sections 5-6)**
     基于神经网络的机器翻译建模，包括：张量计算基础、基于神经网络的语言模型和翻译模型

说了半天，机器翻译是个啥东西？

- 如果你对自然语言处理技术有了解
  - 那就不需要解释了

vs.

说了半天，机器翻译是个啥东西？

- 如果你对自然语言处理技术有了解
  - 那就不需要解释了

vs.

- 如果你使用过Baidu翻译且对软件技术有些了解
  - 机器翻译系统实际就是一个软件，输入是待翻译句子/文本，输出是译文句子/文本

说了半天，机器翻译是个啥东西？

- 如果你对自然语言处理技术有了解
  - 那就不需要解释了

 vs. 

- 如果你使用过Baidu翻译且对软件技术有些了解
  - 机器翻译系统实际就是一个软件，输入是待翻译句子/文本，输出是译文句子/文本
- 请对没听说过机器翻译的同学注意：
  - 机器翻译系统和QQ类似，就是可以在计算机上运行的工具
    机器翻译系统 = 软件工具

说了半天，机器翻译是个啥东西？

- 如果你对自然语言处理技术有了解
  - 那就不需要解释了


vs.

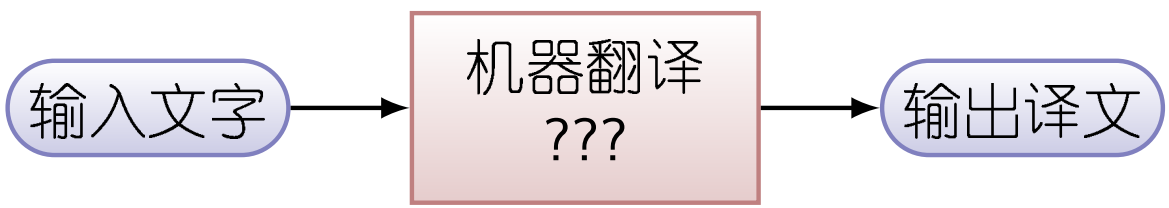

- 如果你使用过Baidu翻译且对软件技术有些了解
  - 机器翻译系统实际就是一个软件，输入是待翻译句子/文本，输出是译文句子/文本
- 请对没听说过机器翻译的同学注意：
  - 机器翻译系统和QQ类似，就是可以在计算机上运行的工具
    机器翻译系统 = 软件工具
  - 但是，机器翻译系统不是QQ
    机器翻译系统 ≠ 好看的按钮
    机器翻译系统 ≠ 美丽的图标

说了半天，机器翻译是个啥东西？

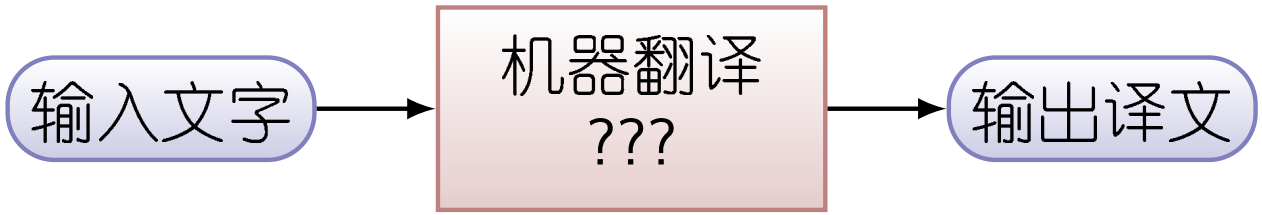
- 如果你对自然语言处理技术有了解
  - 那就不需要解释了

 vs. 

- 如果你使用过Baidu翻译且对软件技术有些了解
  - 机器翻译系统实际就是一个软件，输入是待翻译句子/文本，输出是译文句子/文本
- 请对没听说过机器翻译的同学注意：
  - 机器翻译系统和QQ类似，就是可以在计算机上运行的工具
    机器翻译系统 = 软件工具
  - 但是，机器翻译系统不是QQ
    机器翻译系统 ≠ 好看的按钮
    机器翻译系统 ≠ 美丽的图标
  - 机器翻译是完成自动翻译所执行的"不可见的程序"



又说了半天，机器翻译到底是个啥东西？

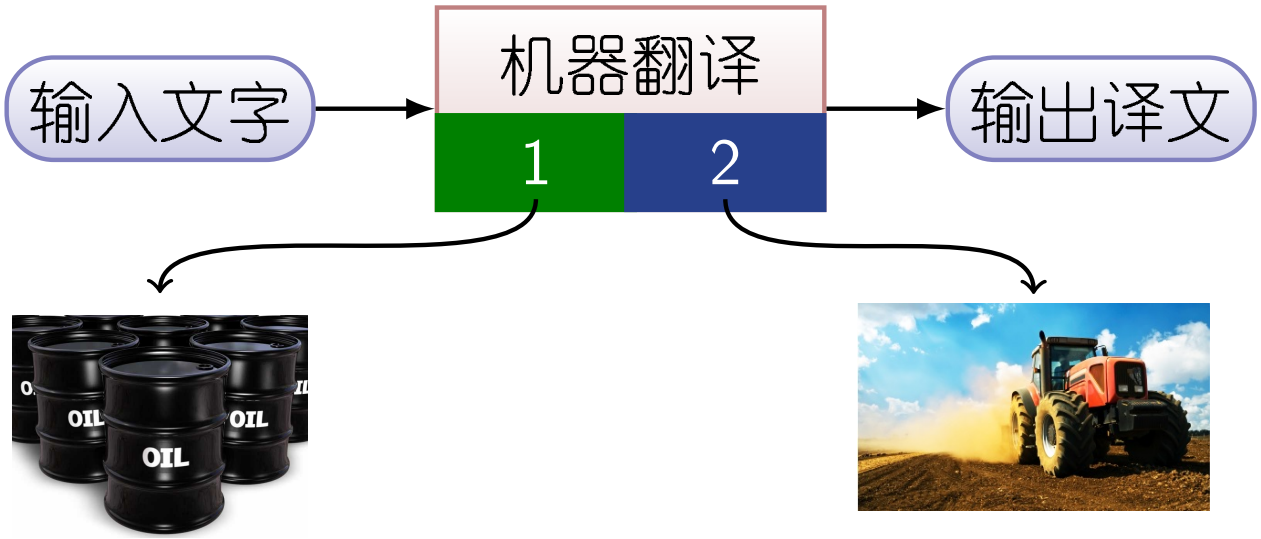


又说了半天，机器翻译到底是个啥东西？

又说了半天，机器翻译到底是个啥东西？

资源：可以使机器翻译系统运行的"汽油"

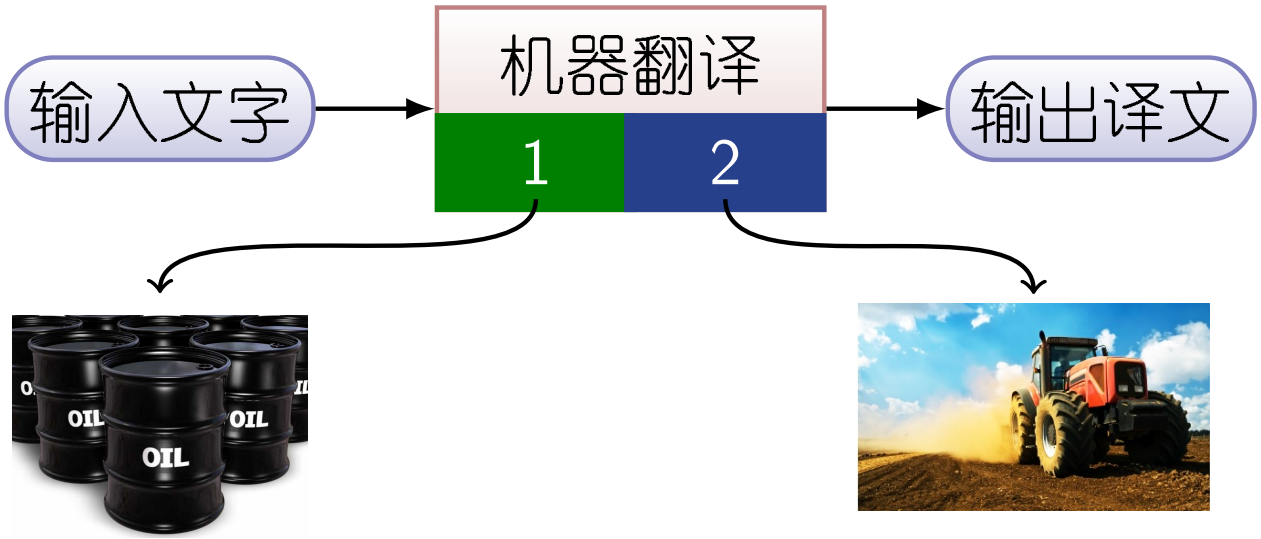
又说了半天，机器翻译到底是个啥东西？



资源：可以使机器翻译系统运行的"汽油"

系统：利用资源完成自动翻译的程序

又说了半天，机器翻译到底是个啥东西？

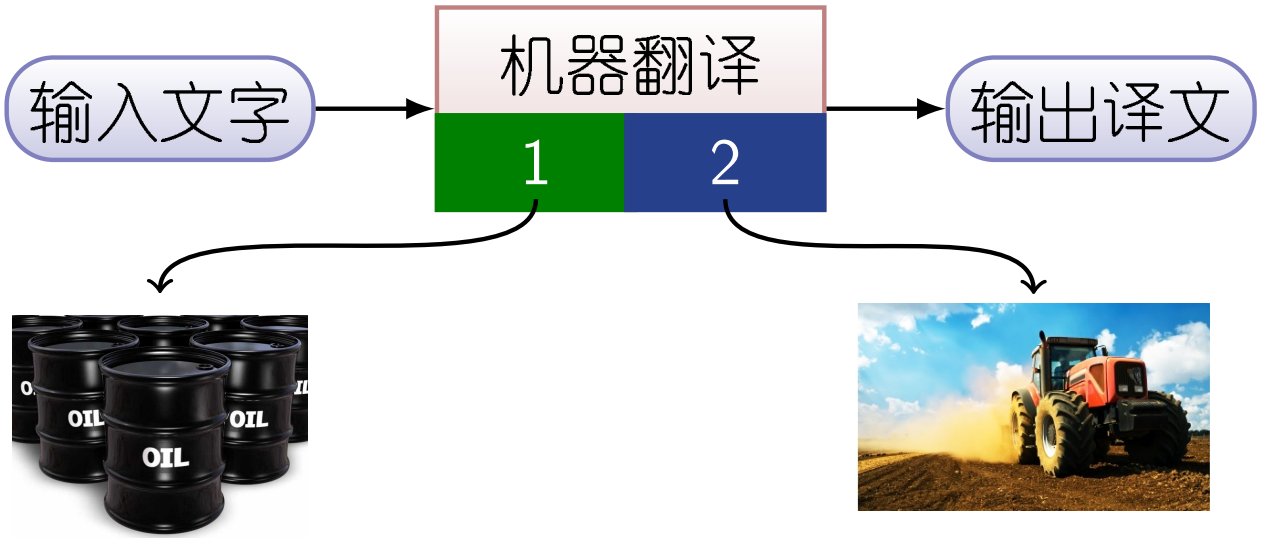


资源：可以使机器翻译系统运行的"汽油"

系统：利用资源完成自动翻译的程序

- 可以说，所有机器翻译系统都是由两部分组成
  1. 资源是指翻译规则、双(单)语数据、知识库等等翻译知识，且这些'知识'是计算机可读的
  2. 系统是各种机器翻译算法的程序实现

又说了半天，机器翻译到底是个啥东西？



资源：可以使机器翻译系统运行的"汽油"

系统：利用资源完成自动翻译的程序

- 可以说，所有机器翻译系统都是由两部分组成
  1. 资源是指翻译规则、双(单)语数据、知识库等等翻译知识，且这些'知识'是计算机可读的
  2. 系统是各种机器翻译算法的程序实现

注意：任何机器翻译系统没有翻译资源的支持都是白搭！

如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构

如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构

我 对 你 感到 满意

如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

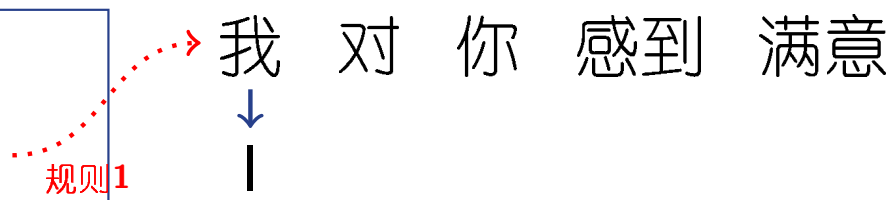
2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

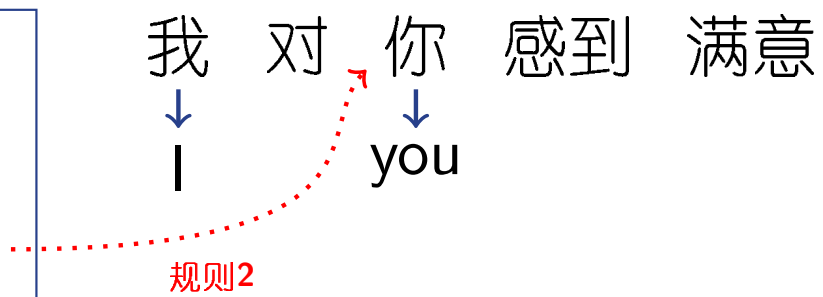
2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

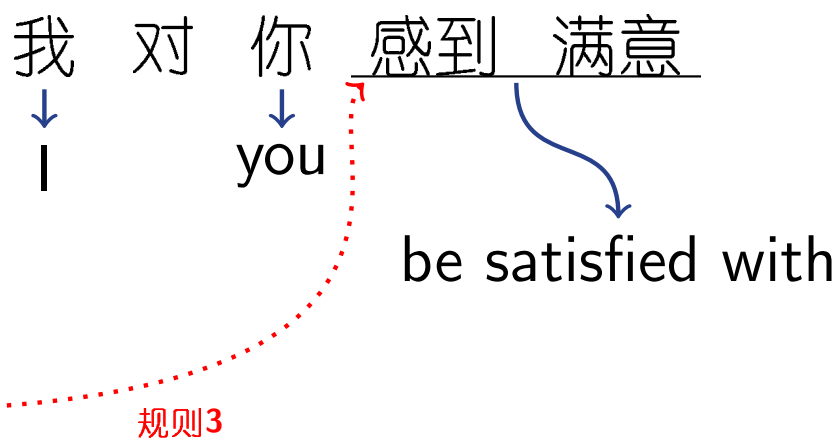
2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

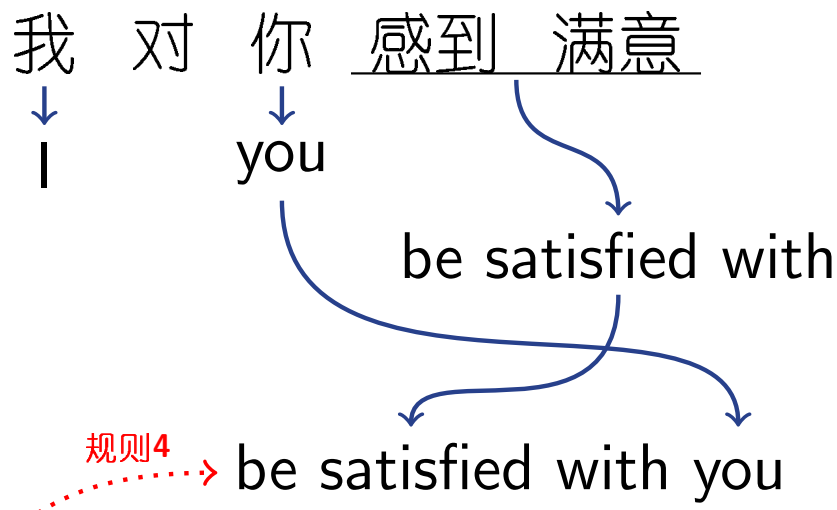
2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构

如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: If 源='我', **then** 译='I'

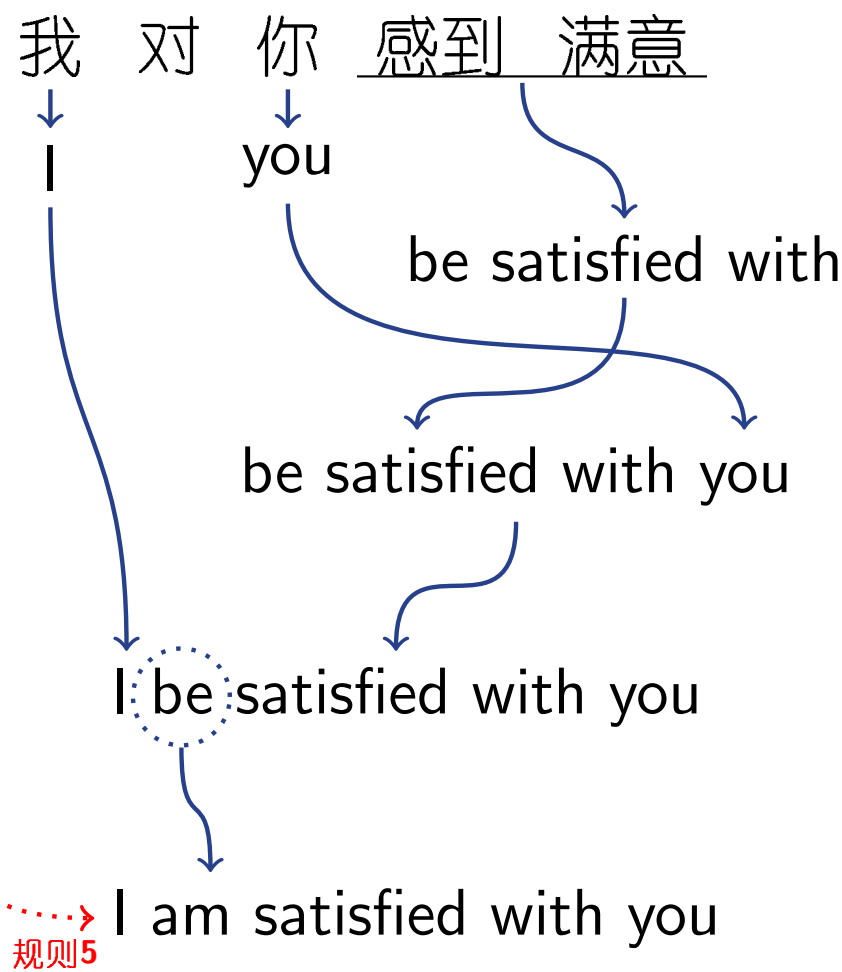
2: If 源='你', **then** 译='you'

3: If 源='感到 满意',
then 译='be satisfied with'

4: If 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: If 译文主语是'I'
then be动词为'am/was'

6: If 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于规则的方法

基于规则的机器翻译：以词典和（人工书写的）规则库构成知识源，用一系列规则的组合完成翻译

- 规则的形式：**If** ..., **then** ...

资源：规则库

1: **If** 源='我', **then** 译='I'

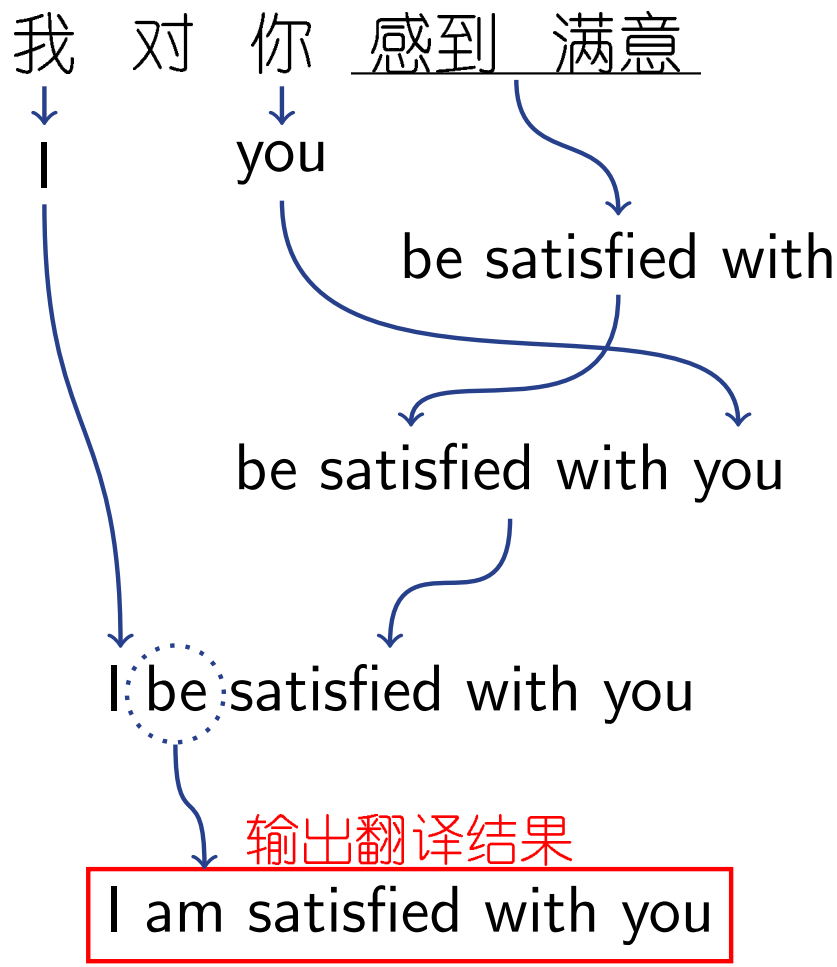
2: **If** 源='你', **then** 译='you'

3: **If** 源='感到 满意',
then 译='be satisfied with'

4: **If** 源='对 ... 动词[表态度]'
then 调序[动词 + 对象]

5: **If** 译文主语是'I'
then be动词为'am/was'

6: **If** 源语是主谓结构
then 译文为主谓结构



如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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 ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

我 对 你 感到 满意

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 ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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

相似度=0

资源2：翻译词典

1: 我 → I | me

2: 你 → you

3: 满意 → satisfy | satisfied ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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

相似度=0.6

资源2：翻译词典

1: 我 → I | me

2: 你 → you

3: 满意 → satisfy | satisfied ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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 ...

...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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 ...
...

如何进行机器翻译？ - 基于实例的方法

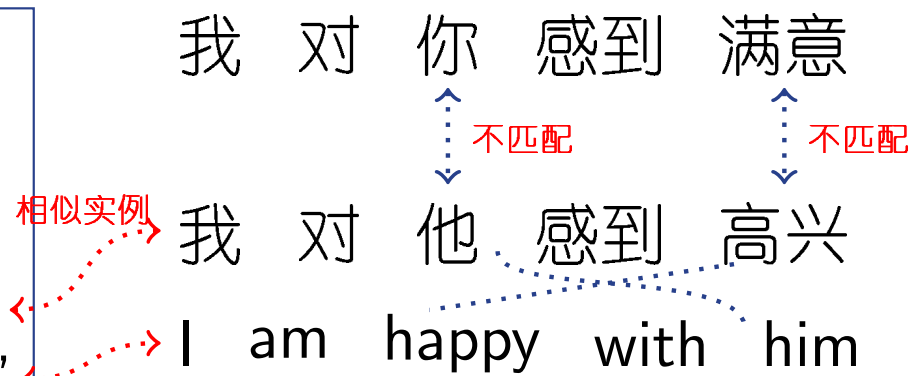
基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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 ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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 ...
...

如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

资源1：翻译实例库

1: 源='什么 时候 开始 ?'
译='When will it start ?'

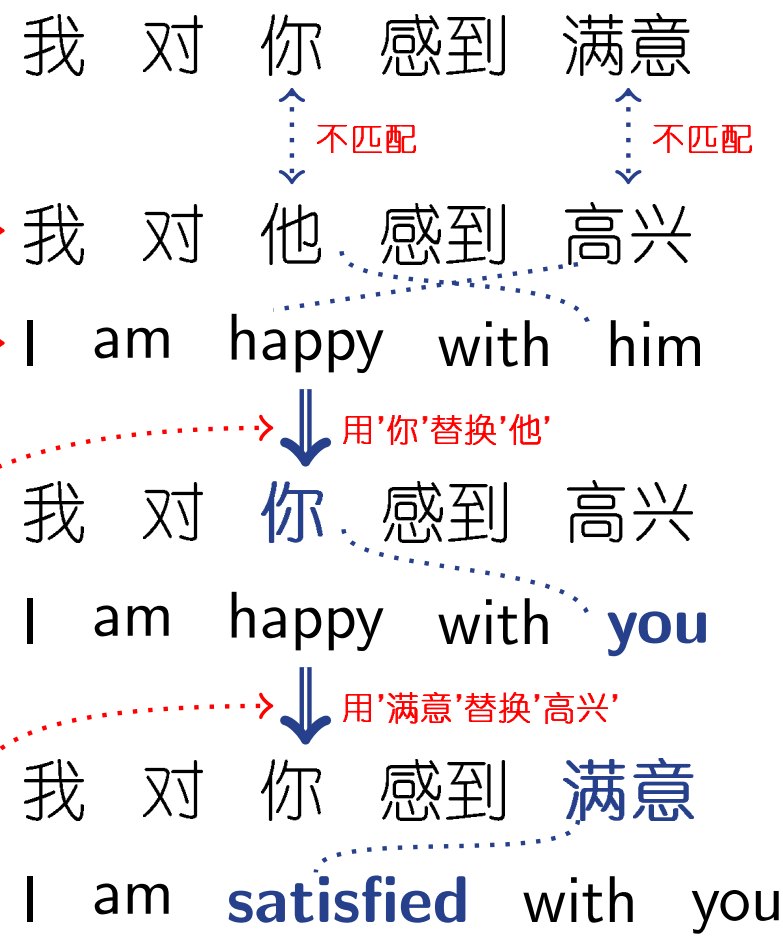
2: 源='我 对 他 感到 高兴'
译='I am happy with him'
...

资源2：翻译词典

1: 我 → I | me

2: 你 → you

3: 满意 → satisfy | satisfied
...



如何进行机器翻译？ - 基于实例的方法

基于实例的机器翻译：在双语句库中找到与待翻译句子相似的实例，之后对其译文进行必要修改，得到最终的译文

- 特例：翻译记忆(TM)只查找能够精确匹配的例句输出

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
...

如何进行机器翻译？ - 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

如何进行机器翻译？ - 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

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!'
...

如何进行机器翻译？ - 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

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

$\Pr(\text{我} \rightarrow \text{I}) = 0.7$

$\Pr(\text{我} \rightarrow \text{me}) = 0.3$

$\Pr(\text{你} \rightarrow \text{you}) = 0.9$

$\Pr(\text{开心} \rightarrow \text{happy}) = 0.5$

$\Pr(\text{满意} \rightarrow \text{satisfied}) = 0.4$

...

如何进行机器翻译？ - 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

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 NiuTrans ?

2: Are you fulfilled ?

3: Yes, you are right .
...

如何进行机器翻译？ - 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

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 NiuTrans ?

2: Are you fulfilled ?

3: Yes, you are right .
...

学习 →

语言模型

$\Pr(\text{I}) = 0.0001$

$\Pr(\text{I} \rightarrow \text{am}) = 0.623$

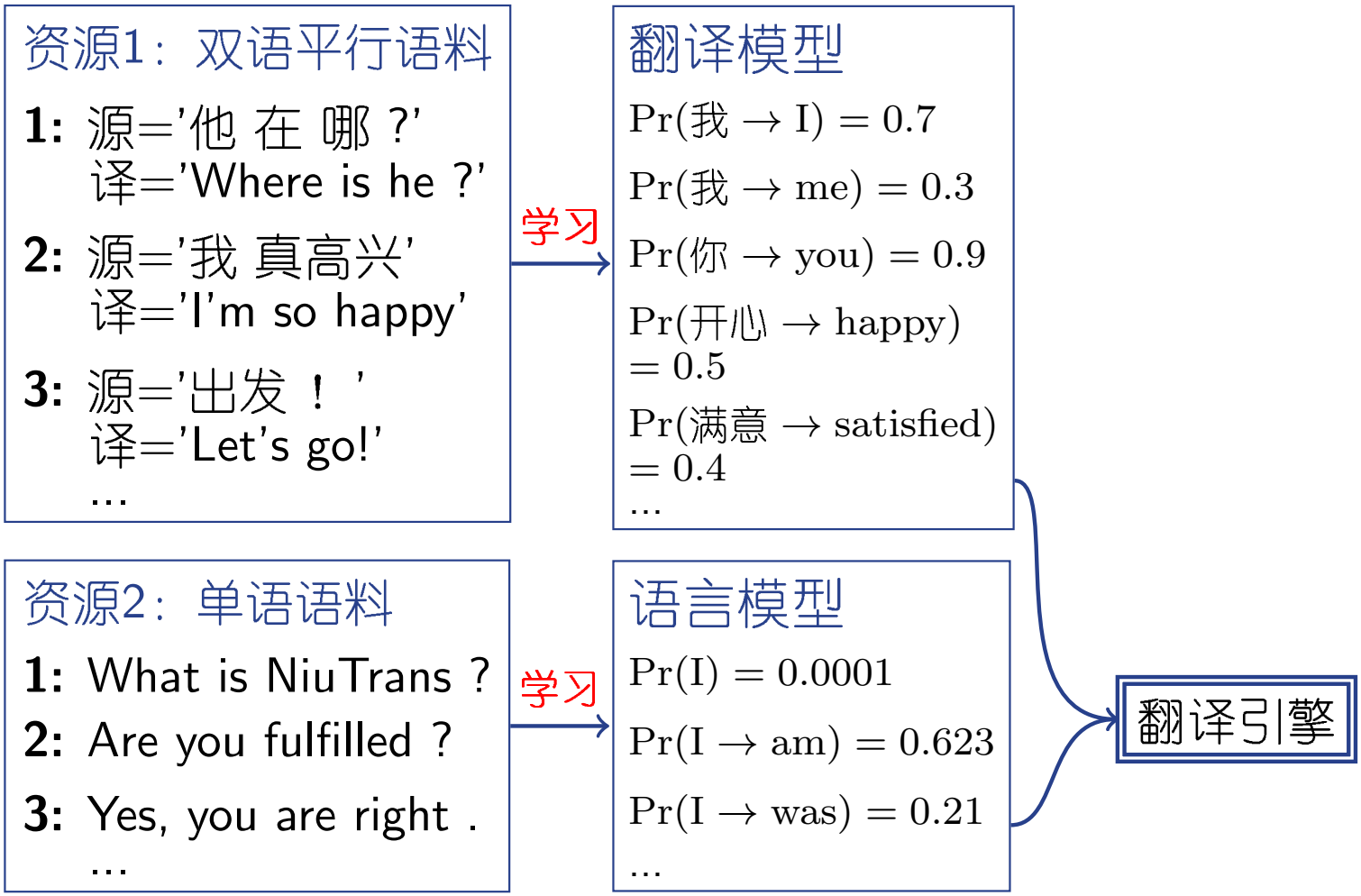
$\Pr(\text{I} \rightarrow \text{was}) = 0.21$

...

如何进行机器翻译？ - 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

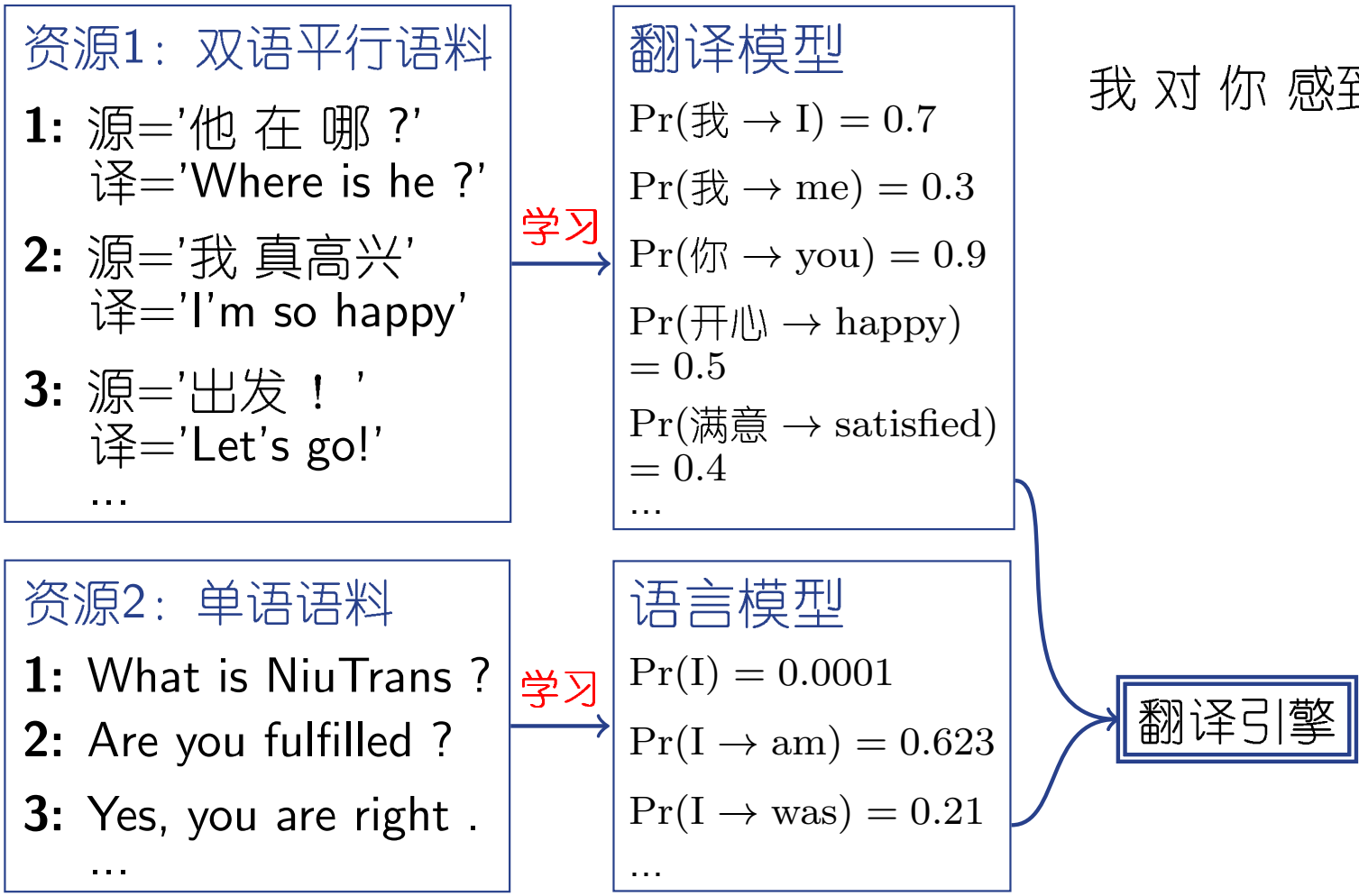
- 建模 + 学习知识 + 利用这些知识进行翻译



如何进行机器翻译？ - 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

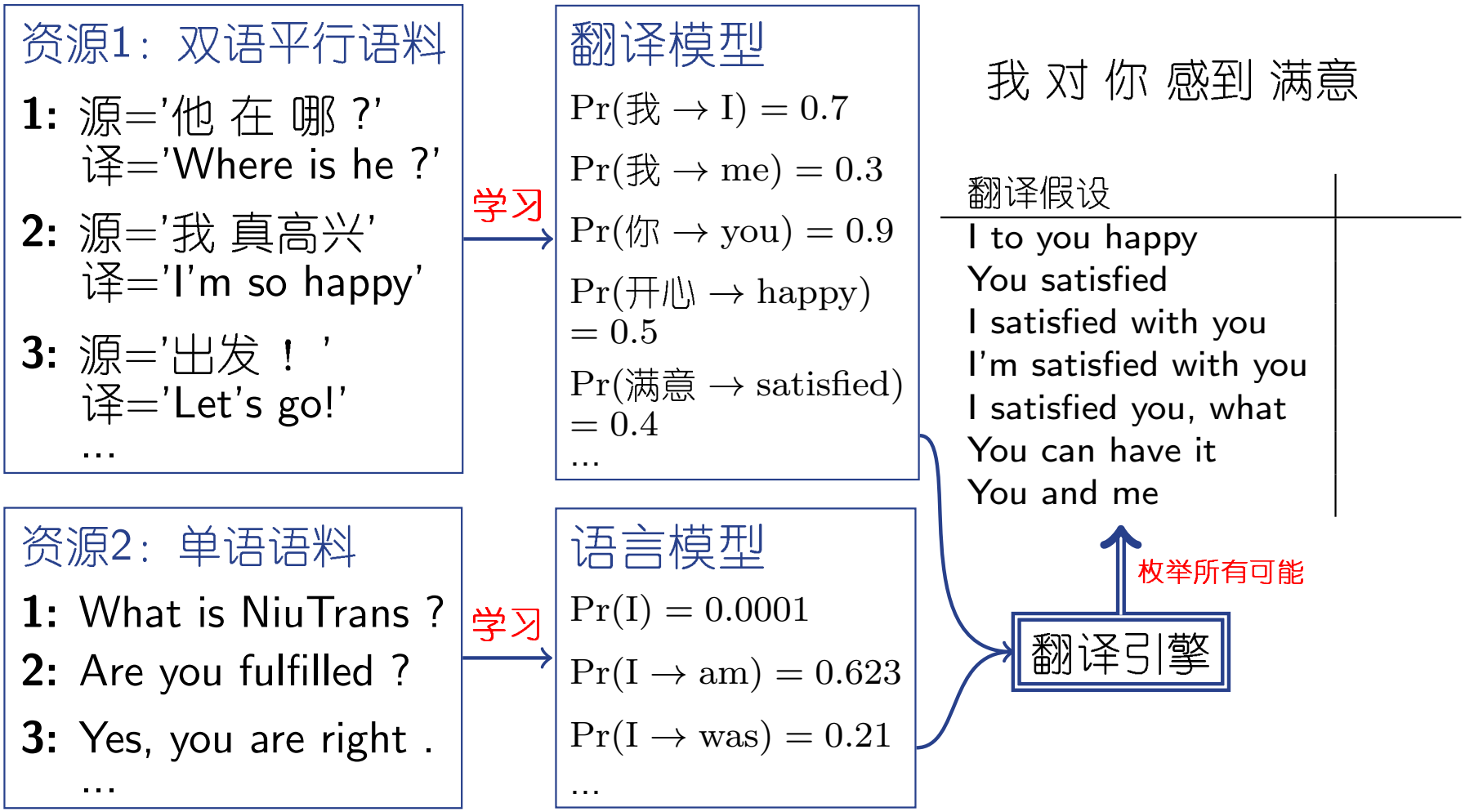
- 建模 + 学习知识 + 利用这些知识进行翻译



如何进行机器翻译？ - 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

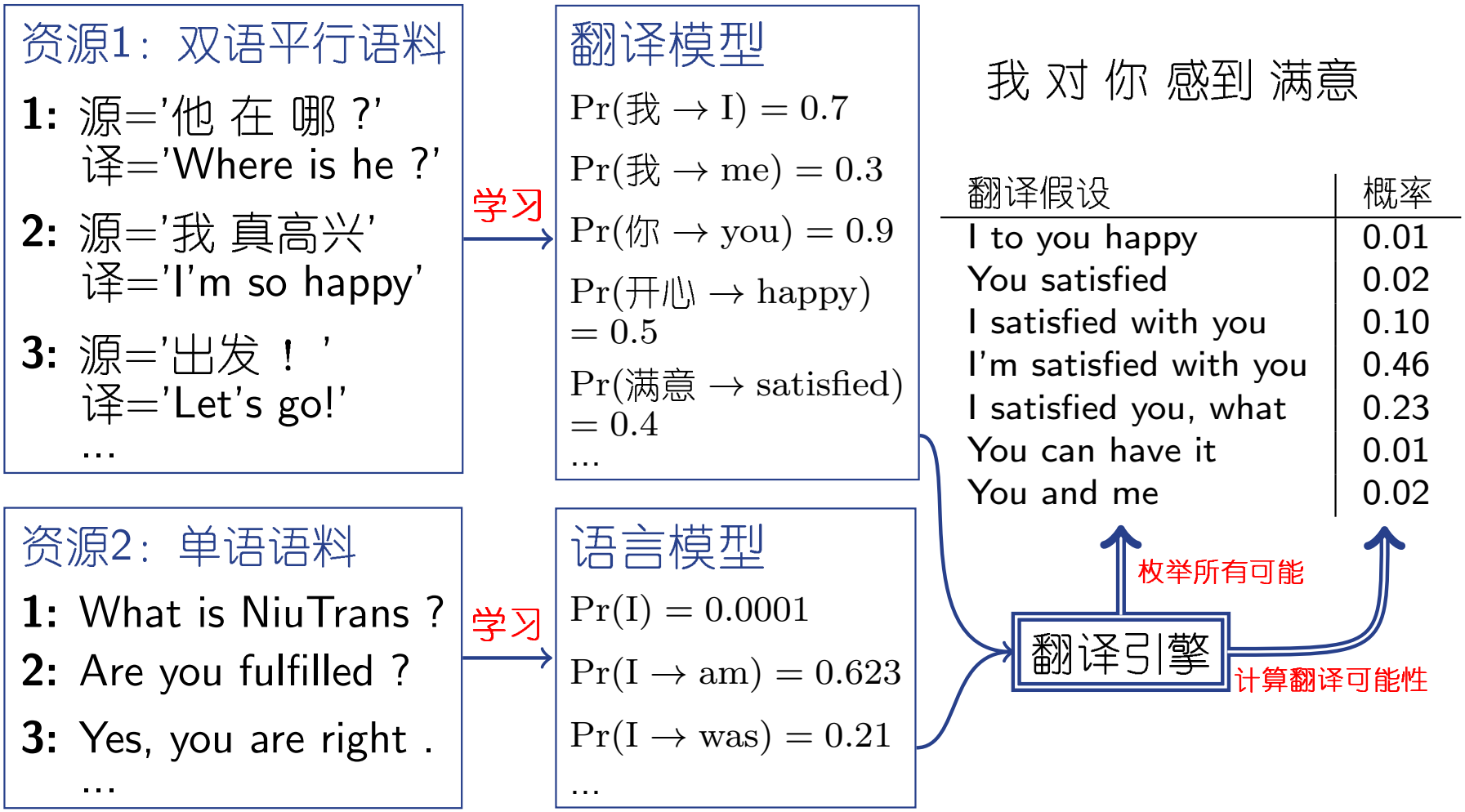
- 建模 + 学习知识 + 利用这些知识进行翻译



如何进行机器翻译？ - 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

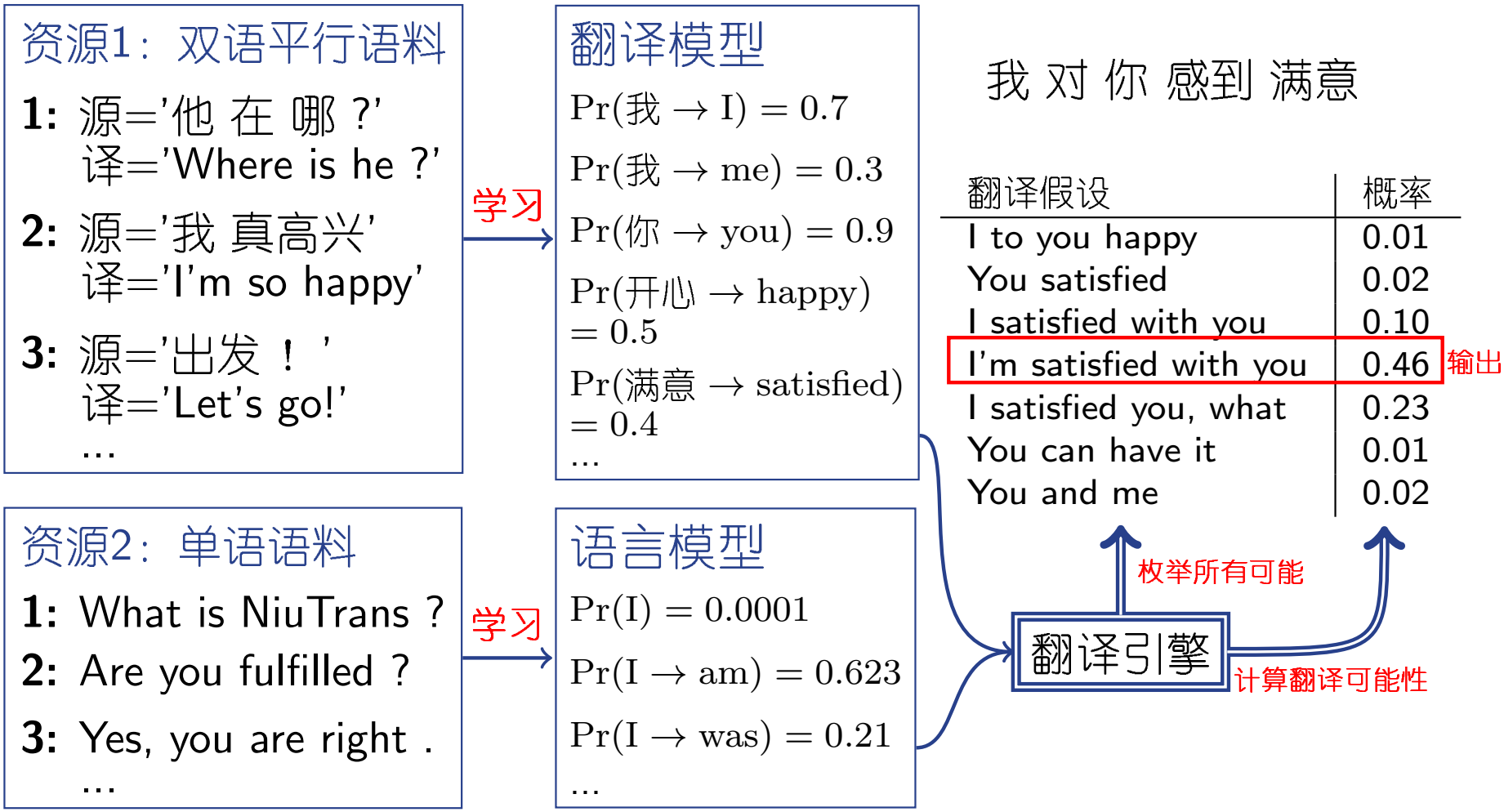
- 建模 + 学习知识 + 利用这些知识进行翻译



如何进行机器翻译？ - 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



如何进行机器翻译？ - 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

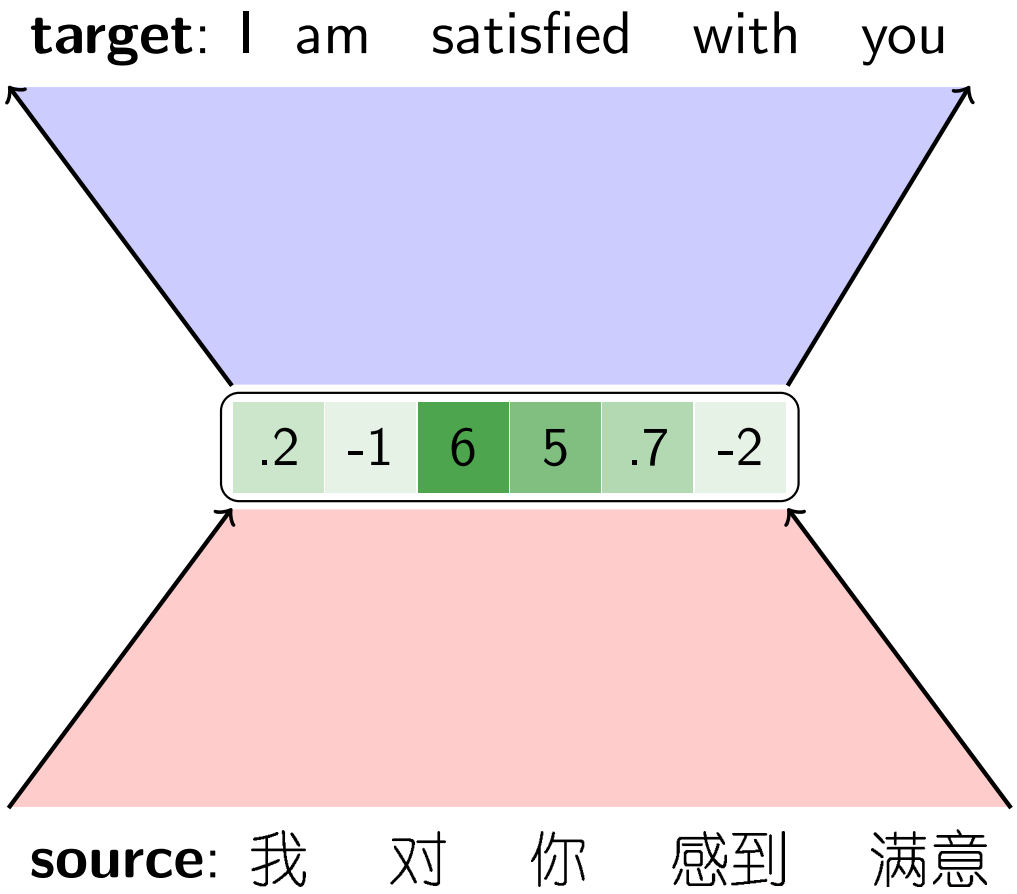
source: 我　对　你　感到　满意

如何进行机器翻译？ - 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

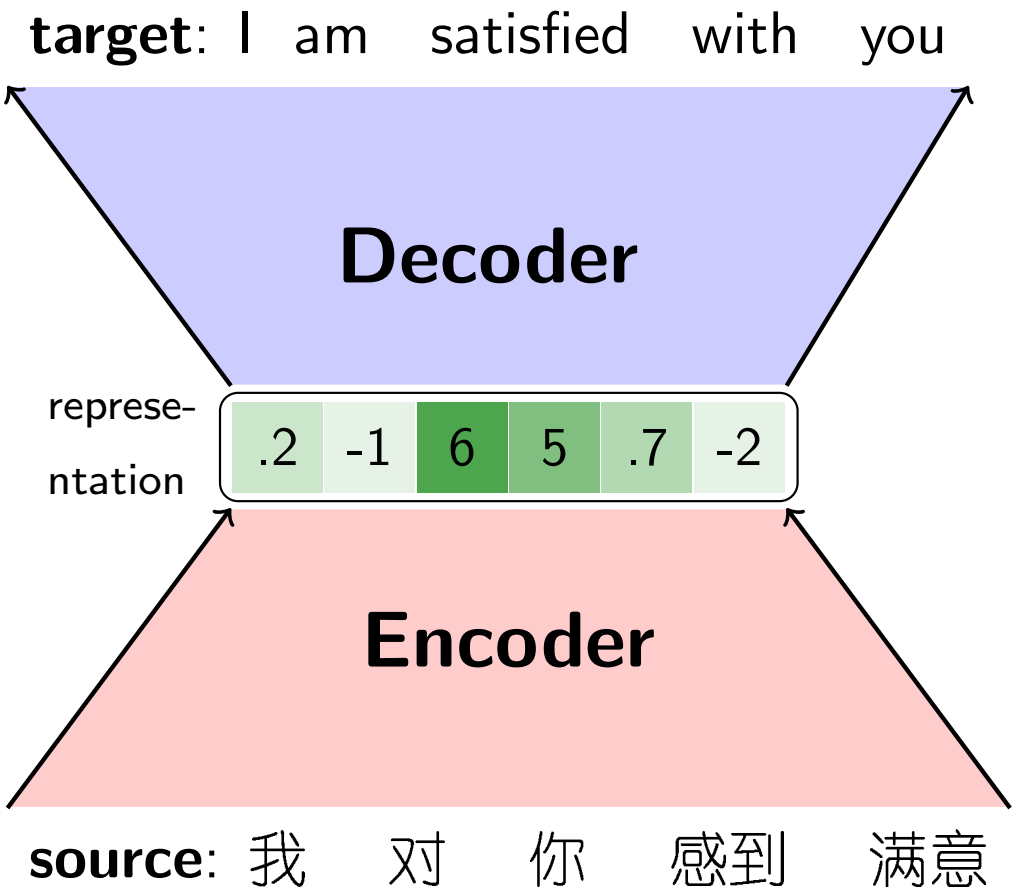
如何进行机器翻译？ - 基于神经网络的方法

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



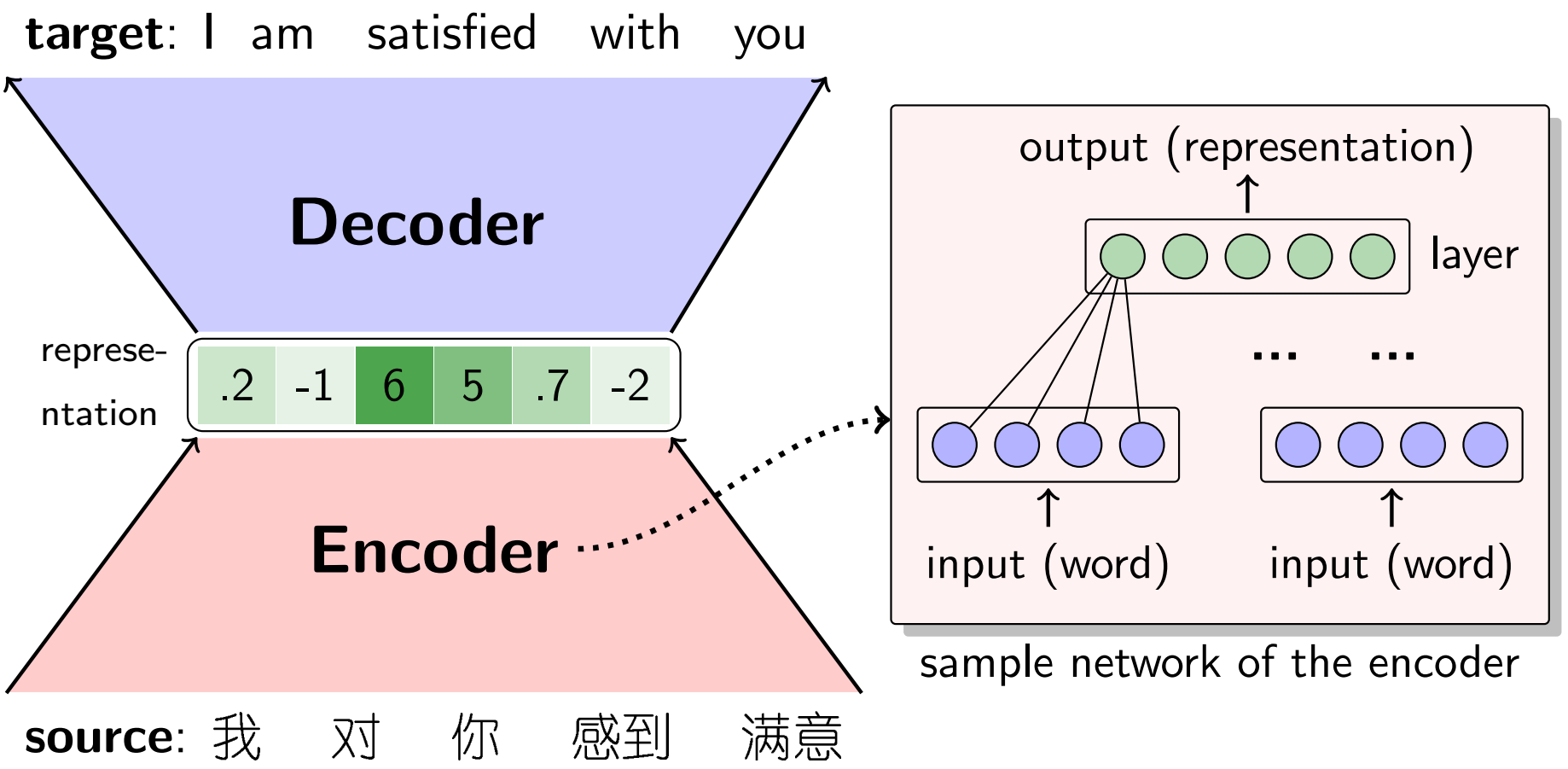
如何进行机器翻译？ - 基于神经网络的方法

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



如何进行机器翻译？ - 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



对比

不同机器翻译方法

	规则	实例	统计	神经
人工写规则	是	否	否	否
人工代价	高	一般	几乎无	几乎无

对比

不同机器翻译方法

	规则	实例	统计	神经
人工写规则	是	否	否	否
人工代价	高	一般	几乎无	几乎无
数据驱动	否	是	是	是
依赖数据质量	N/A	高	低	较低
抗噪声能力	低	低	高	较高

对比

不同机器翻译方法

	规则	实例	统计	神经
人工写规则	是	否	否	否
人工代价	高	一般	几乎无	几乎无
数据驱动	否	是	是	是
依赖数据质量	N/A	高	低	较低
抗噪声能力	低	低	高	较高
使用范围	受限	受限	通用	通用
翻译质量	较高(受限)	较高(受限)	高	很高

对比

不同机器翻译方法

	规则	实例	统计	神经
人工写规则	是	否	否	否
人工代价	高	一般	几乎无	几乎无
数据驱动	否	是	是	是
依赖数据质量	N/A	高	低	较低
抗噪声能力	低	低	高	较高
使用范围	受限	受限	通用	通用
翻译质量	较高(受限)	较高(受限)	高	很高

- 基于规则和实例的机器翻译是傻子：依赖一定人工，规则和模板能够匹配的情况下翻译质量高，但是系统泛化能力有限
- 统计和神经机器翻译是疯子：只依赖数据，不依赖人工，系统健壮性强，但是精度不稳定且翻译过程难以人工干预

学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译

学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译
- 如果上面的理由还不够充分，我们可以换个说法

学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译
- 如果上面的理由还不够充分，我们可以换个说法
  - MT是很cool的technology - 什么big data、什么deep learning都可以往上扯

学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译
- 如果上面的理由还不够充分，我们可以换个说法
  - MT是很cool的technology - 什么big data、什么deep learning都可以往上扯
  - MT能帮你找个好job - 看看最近互联网公司的招聘广告？
  - MT是你paper work的很好的题目 - 看看ACL的论文
  - MT是你拿project非常好的选题 - 看看做SMT的老师

学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译
- 如果上面的理由还不够充分，我们可以换个说法
  - MT是很cool的technology - 什么big data、什么deep learning都可以往上扯
  - MT能帮你找个好job - 看看最近互联网公司的招聘广告？
  - MT是你paper work的很好的题目 - 看看ACL的论文
  - MT是你拿project非常好的选题 - 看看做SMT的老师
  - MT能帮你找个girlfriend - 这个还得靠自己

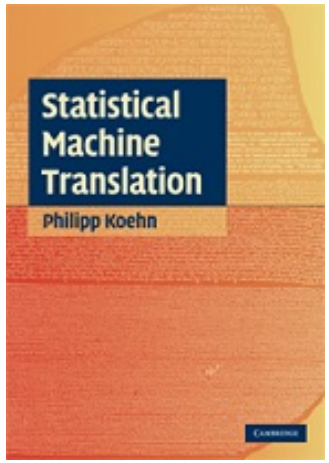
学习机器翻译 - 理由

- 为啥要学习机器翻译(MT)？
  - 当今自然语言处理乃至人工智能领域的热点
  - 应用机器学习框架，数据驱动，不需人工
  - 技术受到广泛关注，已经商用，如Google翻译、Baidu翻译
- 如果上面的理由还不够充分，我们可以换个说法
  - MT是很cool的technology - 什么big data、什么deep learning都可以往上扯
  - MT能帮你找个好job - 看看最近互联网公司的招聘广告？
  - MT是你paper work的很好的题目 - 看看ACL的论文
  - MT是你拿project非常好的选题 - 看看做SMT的老师
  - MT能帮你找个girlfriend - 这个还得靠自己

总之，机器翻译可以帮助你成家立业、改变世界！

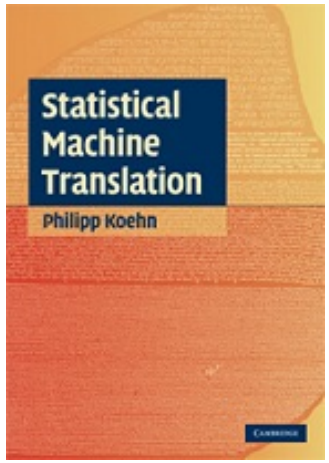
学习机器翻译 - 参考资料

- 推荐教材《Statistical Machine Translation》
  - 经典之作，内容全面
  - 作者是SMT大牛Philipp Koehn
  - 被很多欧美高校和机器翻译研究者使用
  - 建议阅读英文版本，不过也有中文版本



学习机器翻译 - 参考资料

- 推荐教材《Statistical Machine Translation》
  - 经典之作，内容全面
  - 作者是SMT大牛Philipp Koehn
  - 被很多欧美高校和机器翻译研究者使用
  - 建议阅读英文版本，不过也有中文版本



- 一个介绍机器翻译研究的网站
  `http://www.statmt.org/`
- 关于机器翻译和自然语言处理的论文
  `http://www.mt-archive.info/`
  `http://anthology.aclweb.org/`
- 关注一些会议，
  如ACL、NAACL、EMNLP、COLING、WMT、AMTA
- Google搜索一下machine translation course

学习机器翻译 - 开源机器翻译系统

机器翻译的发展可以归功于两个重要方面：

开源系统 + 评测

学习机器翻译 - 开源机器翻译系统

机器翻译的发展可以归功于两个重要方面：
开源系统 + 评测

- **Moses**
  `http://www.statmt.org/moses/`
  - 做MT的人都知道
  - Philipp Koehn所带领团队开发
  - 历史较长、性能优良、功能最全面

学习机器翻译 - 开源机器翻译系统

机器翻译的发展可以归功于两个重要方面：
开源系统 + 评测

- **Moses**
  `http://www.statmt.org/moses/`
  - 做MT的人都知道
  - Philipp Koehn所带领团队开发
  - 历史较长、性能优良、功能最全面

- **NiuTrans**
  `http://www.nlplab.com/NiuPlan/NiuTrans.html`
  - 东北大学自然语言处理实验室开发
  - 模型完整、性能不差、使用简单
  - 良好的中文技术支持

学习机器翻译 - 开源机器翻译系统

机器翻译的发展可以归功于两个重要方面：
开源系统 + 评测

- **Moses**
  `http://www.statmt.org/moses/`
  - 做MT的人都知道
  - Philipp Koehn所带领团队开发
  - 历史较长、性能优良、功能最全面

- **NiuTrans**
  `http://www.nlplab.com/NiuPlan/NiuTrans.html`
  - 东北大学自然语言处理实验室开发
  - 模型完整、性能不差、使用简单
  - 良好的中文技术支持

- 其它一些非常优秀的开源SMT系统：
  Silkroad、Joshua、cdec、Phrasal、HiFST、Jane等

学习机器翻译 - 开源机器翻译系统(2)

- **Tensor2Tensor**
  https://github.com/tensorflow/tensor2tensor
  - 最新的基于深度学习的系统
  - 支持Transformer等模型
  - 使用Tensorflow，功能强大

- **Fairseq**
  https://github.com/facebookresearch/fairseq
  - Facebook出品，代码易读
  - 基于Pytorch，简单易用

学习机器翻译 - 开源机器翻译系统(2)

- **Tensor2Tensor**
  https://github.com/tensorflow/tensor2tensor
  - 最新的基于深度学习的系统
  - 支持Transformer等模型
  - 使用Tensorflow，功能强大

- **Fairseq**
  https://github.com/facebookresearch/fairseq
  - Facebook出品，代码易读
  - 基于Pytorch，简单易用

- 开源机器翻译系统列表
  https://github.com/NiuTrans/MT-paper-lists

学习机器翻译 - 评测及实验数据

- 现在研究使用的绝大多数数据都可以从**LDC**(Linguistic Data Consortium)上找到
  `https://www.ldc.upenn.edu/`

学习机器翻译 - 评测及实验数据

- 现在研究使用的绝大多数数据都可以从**LDC**(Linguistic Data Consortium)上找到
  https://www.ldc.upenn.edu/
- 很多权威机构也组织各种机器翻译评测（比赛），可以作为研究的benchmark
  - **NIST评测**：历史悠久，影响力巨大，翻译任务多，难度大，数据使用最广泛
    http://www.itl.nist.gov/iad/mig/tests/mt/
  - **CWMT评测**：国内评测，发展迅速，翻译语言对丰富
    http://nlp.ict.ac.cn/eval.php
  - **WMT评测**：欧洲语系翻译，数据可以免费下载
    http://www.statmt.org/wmt14/
  - **IWSLT评测**：口语翻译评测
    http://workshop2013.iwslt.org/
  - **NTCIR PatentMT**:专利翻译评测
    http://ntcir.nii.ac.jp/PatentMT-2/

关于课程

关于课程有几点需要注意：

- 机器翻译的学习需要一些程序开发基础，任意的编程语言都可以
  - C/C++，Python，Perl等
  - 没有编程基础咋办？不耽误课程学习，但是练习无法进行
- 语言学知识、数学知识多多益善，但不是必须

关于课程

关于课程有几点需要注意：

- 机器翻译的学习需要一些程序开发基础，任意的编程语言都可以
  - C/C++，Python，Perl等
  - 没有编程基础咋办？不耽误课程学习，但是练习无法进行
- 语言学知识、数学知识多多益善，但不是必须
- 这门课程比较适合计算机相关专业的本科及研究生学习
  - 国内很多高校和研究机构都开展机器翻译的研究
  - 已经有很多相关的课程（如《自然语言处理》）
  - 课件可以参考

关于课程

关于课程有几点需要注意：

- 机器翻译的学习需要一些程序开发基础，任意的编程语言都可以
  - C/C++，Python，Perl等
  - 没有编程基础咋办？不耽误课程学习，但是练习无法进行
- 语言学知识、数学知识多多益善，但不是必须
- 这门课程比较适合计算机相关专业的本科及研究生学习
  - 国内很多高校和研究机构都开展机器翻译的研究
  - 已经有很多相关的课程（如《自然语言处理》）
  - 课件可以参考
- **最后，也是最重要的 -**
  课程的所有练习都以NiuTrans为基础!!!!
  - 完全自主开发，开源系统，便于学习代码
  - NiuTrans团队维护，交流方便

第一节内容就这么结束了！

谢谢！ $\Leftrightarrow$ Thank You!