机器翻译概述

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

什么是机器翻译

- 研究目标：研制出能把一种自然语言（源语言）的文本翻译为另外一种自然语言（目标语言）的文本的计算机软件系统。
- 制造一种机器，让使用不同语言的人无障碍地自由交流，一直是人类的一个梦想。
- 随着国际互联网络的日益普及，网上出现了以各种语言为载体的大量信息，语言障碍问题在新的时代又一次凸显出来，人们比以往任何时候都更迫切需要语言的自动翻译系统。
- 但机器翻译是一个极为困难的研究课题，无论目前对它的需求多么迫切，全自动高质量的机器翻译系统(FAHQMT)仍将是人类一个遥远的梦。

机器翻译的基本方法

- 机器翻译的基本方法
  - 基于规则的机器翻译方法
    - 直接翻译法
    - 转换法
    - 中间语言法
  - 基于语料库的机器翻译方法
    - 基于统计的方法
    - 基于实例的方法
  - 混合式机器翻译方法

目前没有任何一种方法能实现机器翻译的完美理想，但在方法论方面的探索已经使得人们对机器翻译问题的认识更加深刻，而且也确实带动了不少虽不完美但尚可使用的产品问世。

机器翻译的基本方法

- 20世纪90年代以前，机器翻译方法的主流一直是基于规则的方法，因此基于规则的方法也被称为传统的机器翻译方法。
- 直接翻译法
  - 逐词进行翻译，又称逐词翻译法(word for word translation)
  - 无需对源语言文本进行分析
  - 对翻译过程的认识过渡简化，忽视了不同语言之间在词序、词汇、结构等方面的差异。
  - 翻译效果差，属于早期过时认识，现已无人采用

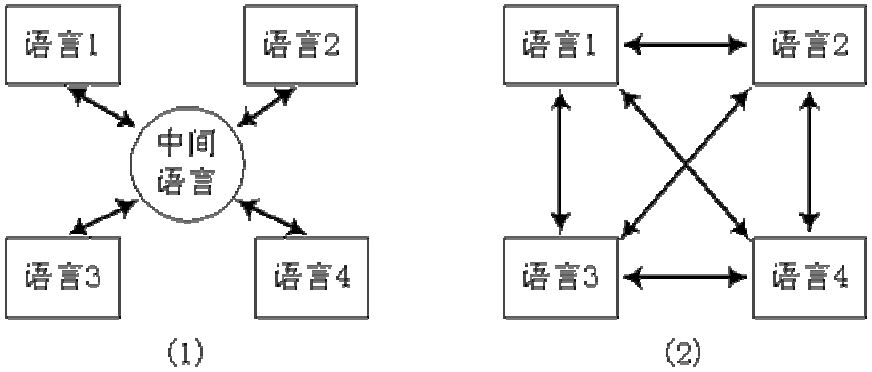
How are you ?　　怎么 是 你 ?
How old are you ?　怎么 老 是 你 ?

机器翻译的基本方法

- 中间语言法(interlingua approach)
  - 中间语言(interlingua)是一种中间表达，通常是一种句法-语义表达(syntactic-semantic expression)，中间语言独立于任何具体的自然语言。
  - 源文本经过深层分析得到其对应的中间语言表示。
  - 再由该中间表示生成目标语文本。
  - 翻译过程为两个阶段。

机器翻译的基本方法

- 不同系统采用不同的中间语言，有的是一种逻辑形式的语言，有的甚至采用类似自然语言的人工语言，如：荷兰政府支持的DLT计划采用世界语Esperanto做中间语言。
- 中间语言法在理论上非常经济，可有效减少翻译模块的数量。可把$n(n\text{-}1)$个直接翻译模块减少为$2n$个翻译模块。



机器翻译的基本方法

- 把任何一种自然语言翻译成为一种独立的中间语言，需要深层次的语言分析和生成技术，目前没有特别成功的基于中间语言的机器翻译系统。

- 转换法(transfer approach)
  - 分析源语言文本，得到源语言的内部表达
  - 将源语言内部表达转换成目标语内部表达
  - 根据目标语内部表达生成目标语文本
  - 翻译过程分成三个阶段

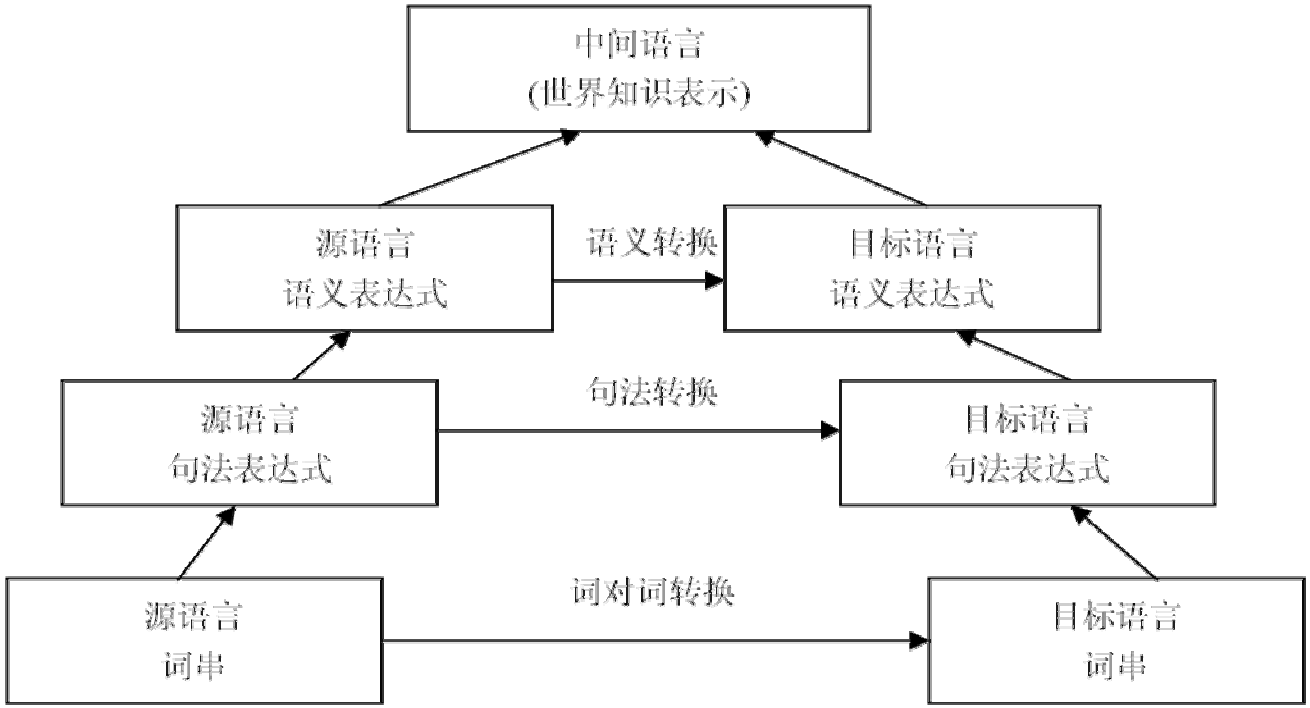
机器翻译的基本方法

- 不同系统采用不同层次内部表示，例如浅层句法表示或深层句法语义表示。
- 商业上最为成功的方法，目前绝大部分商品化机器翻译系统采用转换式机器翻译方法。

- 基于知识的机器翻译方法(knowledge-based machine translation)
  - 20世纪70年代，受人工智能、知识工程发展的影响，而提出。
  - 强调对源语言进行更为彻底的分析和理解。
  - 不仅进行深层语言学分析，还需要进行世界知识(world knowledge)的显式处理。
  - 需要建立对语言理解有益的本体知识库(ontology)。

机器翻译的基本方法

- 研制代价昂贵，没有特别成功的案例。



基于规则的翻译方法图示

机器翻译的基本方法

- 20世纪80年代中后期，基于语料库的机器翻译技术得到越来越多的关注。
  - 试图避开知识库建设的困难
  - 试图回避对源语言进行深层语言分析
  - 翻译知识主要来自双语平行语料库
  - 基于实例的翻译通过模仿实例库中已有的翻译基于类比的策略进行翻译。
  - 基于统计的机翻译通过建立统计翻译模型、训练统计模型进而基于统计模型进行翻译。
- 考虑到这些方法背后的哲学背景，也常把基于规则的方法称为理性主义(rationalism)方法，而把基于语料库的方法称为经验主义(empiricism)方法。

机器翻译的使用

- 总而言之，无论采用何种机器翻译方法，目前的机器翻的译文质量都还远不能令人满意。但并不能说明机器翻译一无是处，机器翻译在许多应用场合已在发挥作用。
- 翻译需求的种类
  - 传播型翻译需求(information dissemination)
    - 希望将自己的信息传播出去
    - 跨国企业的产品说明、技术资料
  - 吸收型翻译需求(information assimilation)
    - 希望了解以自己所不通晓的语言为载体的信息
    - 科学工作者之于科技文献

机器翻译的使用

- 对于信息传播型用户而言
  - 译文质量十分关键，跨国公司的所有技术资料都应准确翻译，不存在妥协的可能。
  - 目前机器翻译似乎难以发挥作用
  - 但跨国公司产品数量有限、领域狭窄，可采用子语言技术以及后编辑(post-edit)技术
  - 机器辅助翻译技术和翻译记忆(translation memory)技术广泛使用
  - 机器翻译也可较好保证术语翻译的一致性

机器翻译的使用

- 对于信息吸收型用户而言
  - 往往面临太多的文献需要浏览，但并非对所有文献都有兴趣
  - 机器可以提供一个初步的翻译，虽不准确，但可传达文献的总体思想，有利于用户定位文献
  - 对于选出的文献，可以聘请专人进行译后编辑或聘请专家进行翻译
  - 同聘请翻译人员相比，机器翻译具有廉价和高速的优势

机器翻译的使用

- 目前机器翻译的价值不在于它可以取代翻译专家，而在于它可在一个完整翻译过程的部分环节中有所贡献。

- 机器翻译的价值也体现在它可以带来翻译生产率的提高和翻译成本的降低这两个方面。

- 互联网时代对机器翻译的新需求

  (1) 网页的翻译

  (2)网络聊天室、技术论坛中用户交谈的实时翻译

  (3)跨语言信息检索(Cross Language Information Retrieval)

  (4)跨语言的信息提取

规则系统中的知识表示

- 开发基于规则的机器翻译系统，首先要设计知识表示系统，将翻译过程中所需要的知识以计算机可以操作的形式表述出来。
- 一般而言，翻译过程往往需要下述一些知识的支撑：

  （一）源语言知识。系统利用源语言知识分析源语言句子，得到源语言句子的结构和意义。

  （二）目标语言知识。系统利用目标语言知识，产生可以接受的目标语言句子。

  （三）源语言到目标语言的对译知识。在基于转换的系统中，系统需要根据各种级别的对应关系来完成源语言到目标语言的转换。最基本的是词之间的对译知识。

规则系统中的知识表示

（四）领域知识和世界知识。利用源语言知识、目标语言知识，在领域知识和世界知识的协助下，可以更好地完成对源语言的理解和目标语言的生成。

（五）有关社会、文化和习俗的知识。在人工翻译中，这些知识也起着重要作用。但鉴于目前的处理水平，几乎没有机器翻译系统把该类知识纳入处理范围。人们目前还缺乏有效方法把这些知识以机器可以操作的方式描述出来。

规则系统中的知识表示

- 词典

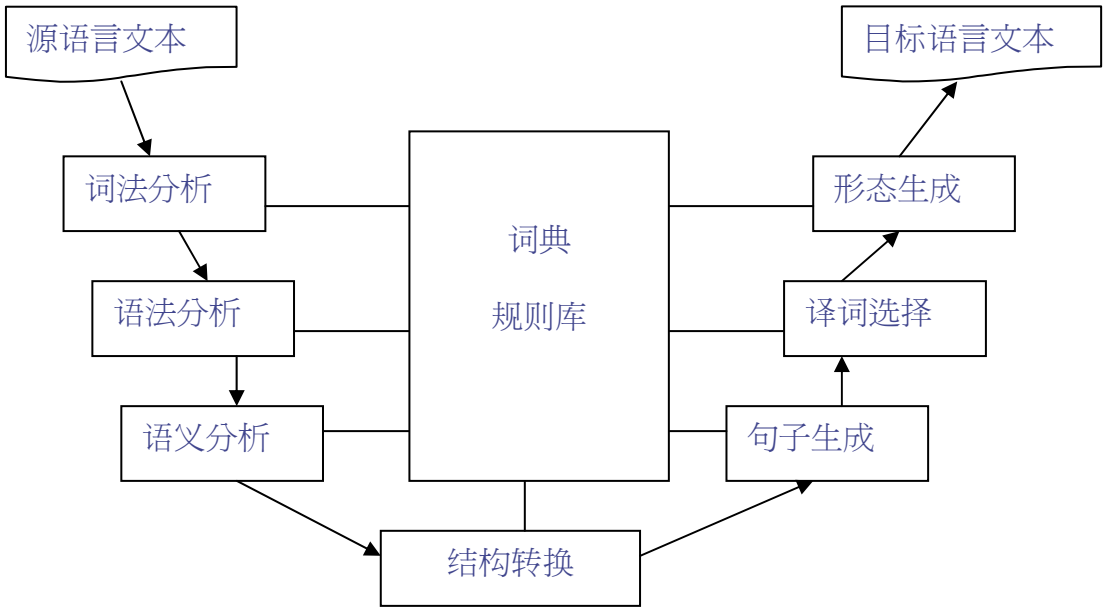
  机器翻译系统中，有关词的知识记录在词典中，源语言的形态知识、句法知识和语义知识记录在**源语言词典**中。目标语言的形态知识、句法知识等记录在**目标语言词典**中。词语间的对译关系则记录在**对译词典**中。

- 规则

  为了源语言句子分析和目标语言句子生成的需要，还需要有关句子结构的知识。句子或短语的组成规律用**规则**描述。源语言和目标语言结构间的对应关系一般用**转换规则**来表达。

规则系统的基本流程

- 词法分析
- 句法分析
- 语义分析
- 结构转换
- 句子生成
- 译词选择
- 形态生成



规则翻译方法例示

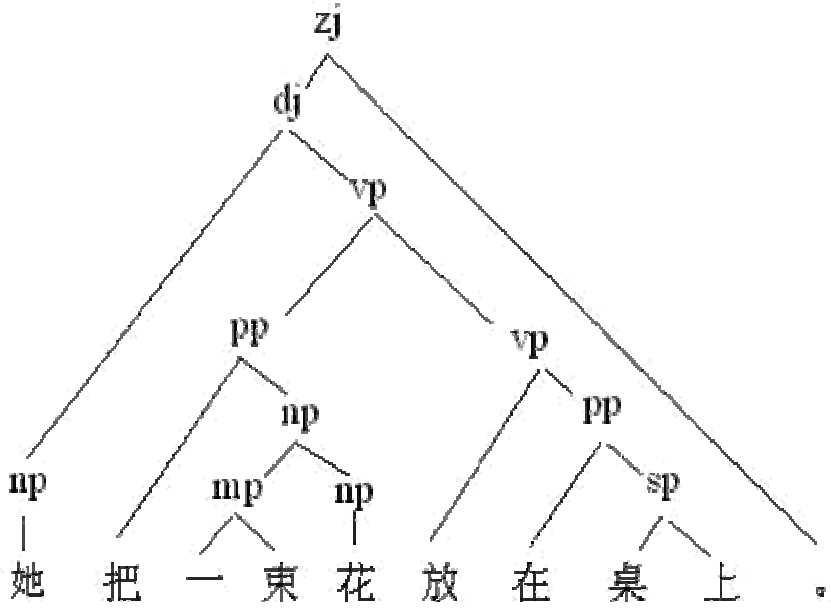
输入：她把一束花放在桌上。

切分和标注：她/r　把/p　一/m　束/q　花/n　放/v　在/p　桌/n　上/f　。/w

规则翻译方法例示

输入：她把一束花放在桌上。

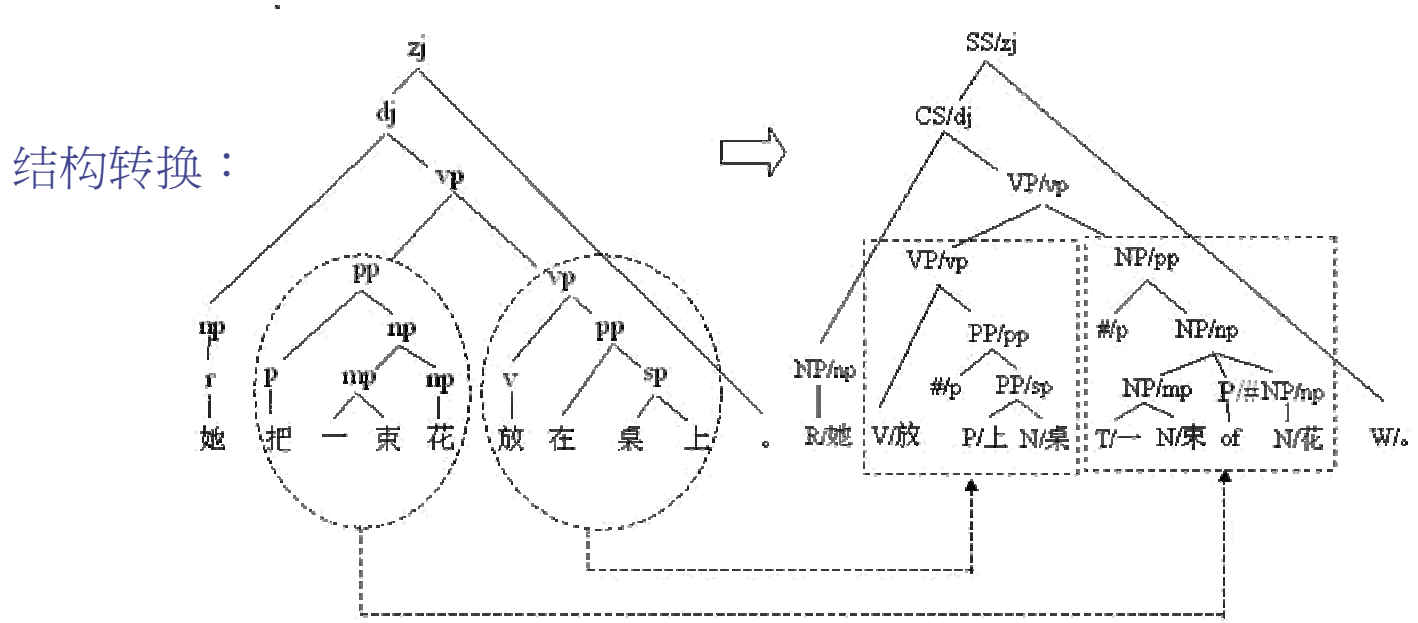
句法分析：



规则翻译方法例示

输入：她把一束花放在桌上。

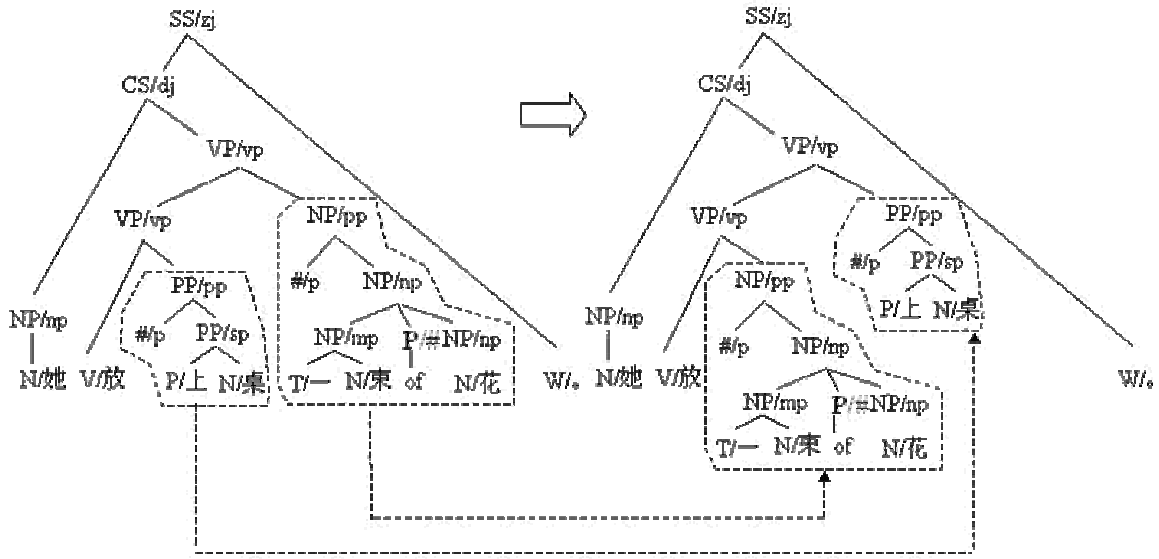
结构转换：



规则翻译方法例示

输入：她把一束花放在桌上。

结构调整：



规则翻译方法例示

输入：她把一束花放在桌上。

译词选择：

她 — she	放 — place	一— a	束 — bunch
花 — flower	上 — on	桌 — table	。 — .

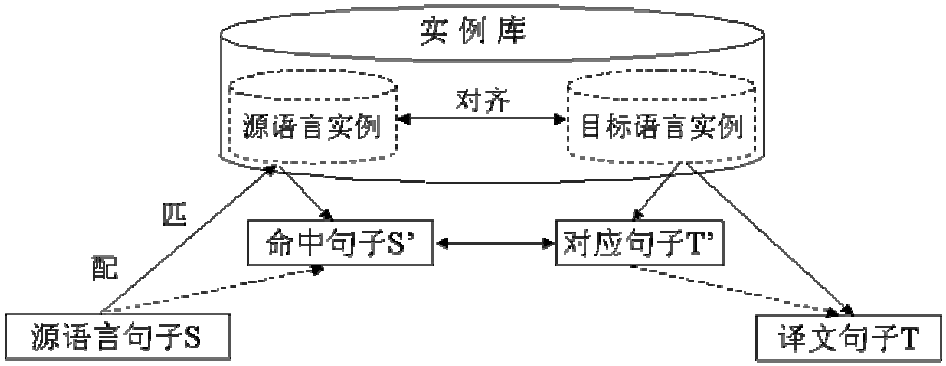
输出：She puts a bunch of flowers on table.

基于实例的机器翻译

- Example Based Machine Translation(EBMT)
- 1984年由日本著名机器翻译专家长尾真提出
- 背景
  - 建立在转换基础上的机器翻译系统，在差异较大语言对间进行互译时，效果不好
  - 人在翻译时不做深层次语言学分析
  - 人在学外语的时候，首先要作大量的对照记忆，在遇到新的句子时，会和记忆中的句子类比
- 建立基于类比思想的机器翻译

基于实例的机器翻译

- 基本思想
  - 主要知识库是双语对照的实例库
  - 当需要翻译一个新句子时，通过检索的办法在实例库中寻找和该句类似的翻译实例。
  - 新句子的翻译可通过模拟最类似的实例的译文的方式获得。



基于实例的机器翻译

- EBMT的优点
  - 系统维护容易
    - 系统中知识以翻译实例和义类词典等形式存在，可以很容易的利用增加实例和词汇的方式扩充系统。
  - 容易产生高质量的译文
    - 尤其是利用了较大的翻译实例或和实例精确匹配时更是如此。
  - 可避免进行深层次的语言学分析
- 类义词典的作用
  - The rabbit eats vegetables
  - Sulfuric acid eats metal
  - He eats apple?

基于实例的机器翻译

- EBMT的关键问题
  - 大规模的双语语料库
  - 双语对齐问题
    - 语篇、句子、短语和词汇等各种级别
  - 建立合理的相似度准则
  - 高效的实例检索机制
  - 译文生成
- 翻译记忆技术和基于模板的翻译技术

基于统计的机器翻译

- Statistic-Based Machine Translation(SBMT)
- 翻译问题是解密问题
- 50年代初曾有提及，遭到以Chomsky为代表的语言学家的反对
- 90年代初，统计翻译技术复苏
  - 统计技术在语音识别领域获得成功
  - 目前计算机性能已能胜任密集型计算
  - 目前也有大量联机双语电子文本

基于统计的机器翻译

- 翻译问题可用噪声信道来描述

  T → 噪音信道 → S

- 基本模型

$$\hat{T} = \arg\max_{T} \Pr(T)\Pr(S \mid T)$$

$$\hat{T} = \arg\max_{T} \Pr(T \mid S)$$

$$\Pr(T \mid S) = \frac{\Pr(T)\Pr(S \mid T)}{\Pr(S)}$$

Pr(*S*|*T*)称为语言*S*到*T*的翻译模型

Pr(*T*) 称为语言*T*的语言模型

基于统计的机器翻译

- 基本问题
  - 建立合理的可计算的语言模型并估计参数
    - n元模型（n-gram）
  - 建立合理的可计算的翻译模型并估计参数
  - 设计可靠有效的算法搜索最好的译文
    - 目前还没有能搜索到最优结果的算法

基于统计的机器翻译

- IBM翻译模型

$$\Pr(S \mid T) = \prod_{i=1}^{n} \left(\Pr(f_i \mid t_i) \cdot \prod_{j=1}^{fi} \Pr(s_j \mid t_i) \right) \cdot \prod_{i,j,l} \Pr(i \mid j,l)$$

$\Pr(f_i \mid t_i)$ 单词 t_i 翻译成 f_i 个单词的概率

$\Pr(s_j \mid t_i)$ 单词 t_i 翻译成单词 s_j 的概率

$\Pr(i \mid j,l)$ 在长度为 l 的源语言句子中，第 i 个位置的单词对应目标语言中第 j 个位置的概率

- 模型训练(EM算法—词对齐)

基于统计的机器翻译

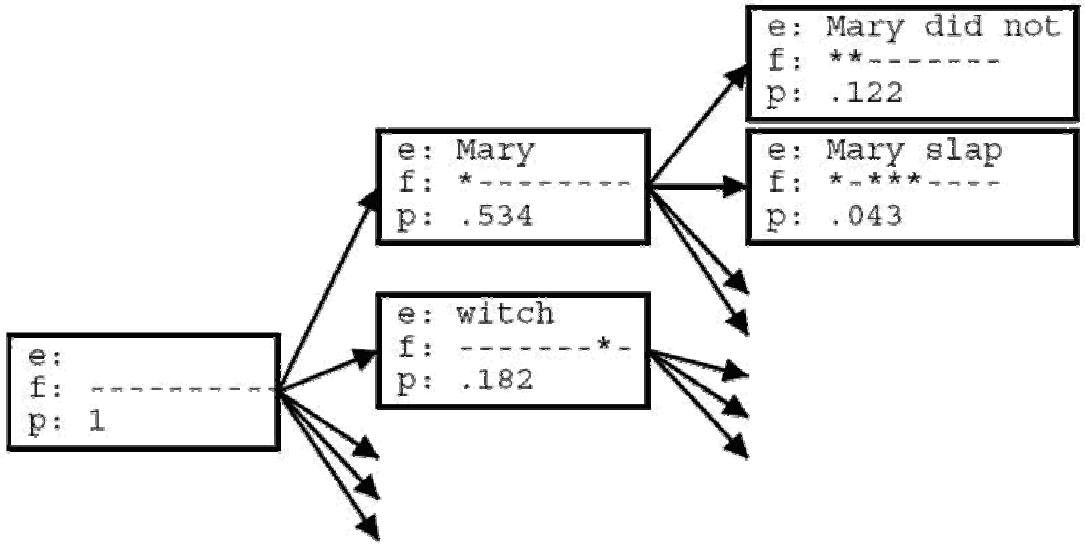
英文: The proposal will not now be implemented

法文: Les(1) propositions(2) ne(4) seront(3) pas(4) mises(7) en(7) application(7) maintenant(5)

$$
\begin{aligned}
&\Pr(1|\textit{The}) \times \Pr(\textit{les}|\textit{the}) \\
&\times \Pr(1|\textit{proposal}) \times \Pr(\textit{propositions}|\textit{proposal}) \\
&\times \Pr(1|\textit{will}) \times \Pr(\textit{seront}|\textit{will}) \\
&\times \Pr(2|\textit{not}) \times \Pr(\textit{ne}|\text{not}) \times \Pr(\textit{pas}|\text{not}) \\
&\times \Pr(1|\textit{now}) \times \Pr(\textit{maintenant}|\textit{now}) \\
&\times \Pr(0|\textit{be}) \\
&\times \Pr(3|\textit{implemented}) \times \Pr(\textit{mises}|\textit{implemented}) \times \Pr(\textit{en}|\textit{implemented}) \times \\
&\quad \Pr(\textit{application}|\textit{implemented}) \\
&\times \Pr(1|1,9) \times \Pr(2|2,9) \times \Pr(3|4,9) \times \Pr(4|3,9) \times \Pr(5|4,9) \times \Pr(6|7,9) \times \Pr(7|7,9) \\
&\quad \Pr(8|7,9) \times \Pr(9|5,9)
\end{aligned}
$$

解码

- Maria no daba una bofetada a la bruja verde (Spanish)
- 穷尽式搜索(exhaustive search)



- 剪枝策略(pruning strategy)

机器翻译为什么困难？

- 语言问题非常复杂，缺乏有效的形式化手段
- 语言中常有大量歧义现象，翻译要面对两种语言间的歧义现象
- 翻译涉及的常是海量知识，知识库的建造维护代价很高
- 机器翻译过程涉及很多环节，每个环节都不能做到100%准确，错误积累严重

机器翻译研究中常用的对策

- 交互式机器翻译
- 子语言（限定领域）
- 受控语言（限定语言的复杂程度）
- 混合式机器翻译

混合式的机器翻译

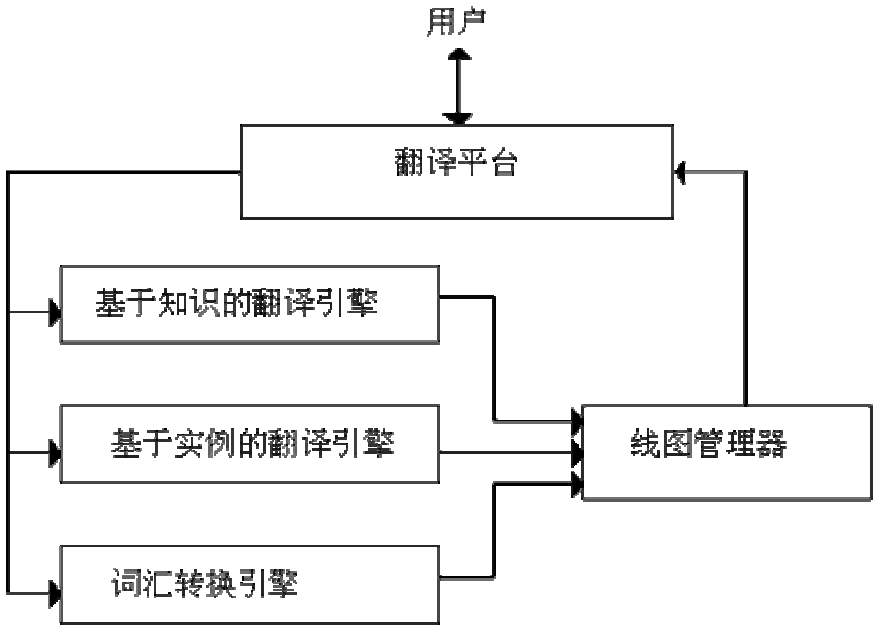
- 20世纪90年代，学界曾就机器翻译方法问题有过争论。
- 坚持规则路线的学者认为
  - 统计模型对结构处理乏力且过于简单？(正在改变)
  - 远距离制约问题？
  - 统计翻译是“石头汤”？
- 坚持统计方法的学者认为
  - 传统方法不能彻底解决机器翻译问题
  - 基于知识的方法曾被认为是解决机器翻译问题的关键方法，可是目前依然没有开发出实用系统，留给大家的是一些写在纸上的例子。

混合式的机器翻译

- 规则方法与统计方法具有互补特点，机器翻译的希望也许在于二者的结合

特点	基于规则的MT	基于统计的MT
健壮性、覆盖范围(robustness/coverage)	不好	较好
质量、流畅性(quality/fluency)	较好	不好
表示(representation)深度	很深	较浅

- 多引擎的机器翻译

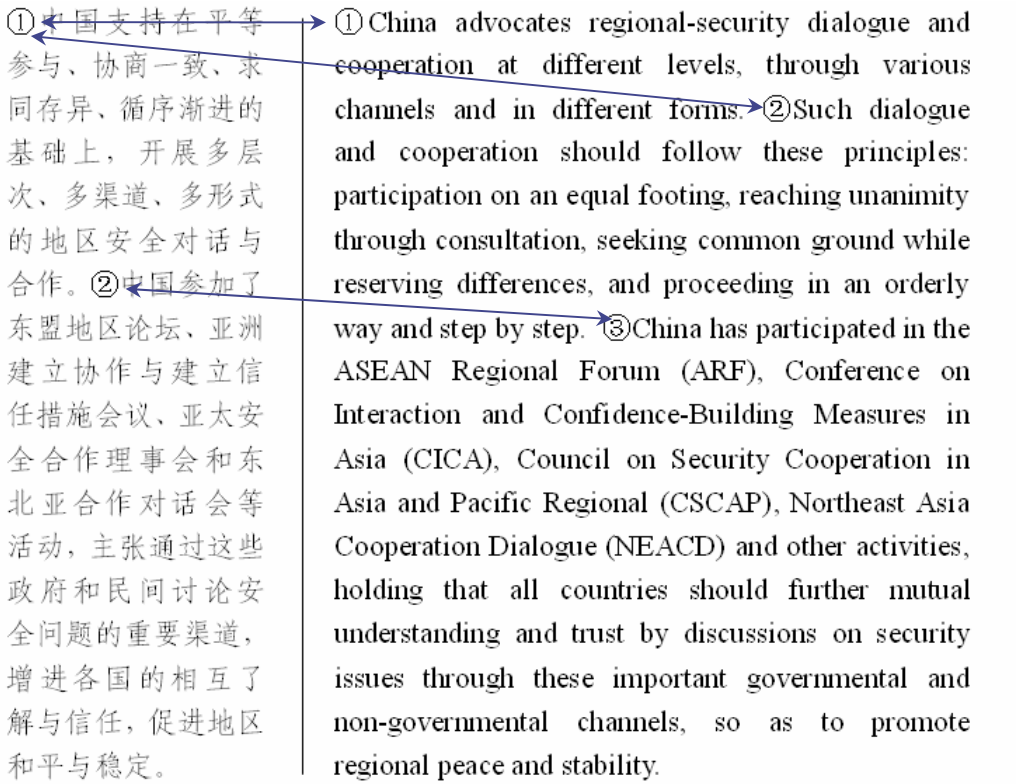


双语语料库和机器翻译

- 大规模双语语料库是机器翻译研究的重要资源
  - 基于双语语料库的翻译知识获取（RBMT）
  - 为基于实例的机器翻译提供翻译实例库
  - 统计机器翻译需要用语料库训练语言模型和统计模型

双语句子级对齐

- 在双语文本间建立句子一级的对齐关系，就是要确定源语言文本中哪个(些)句子和目标语言文本中哪个(些)句子互为译文。

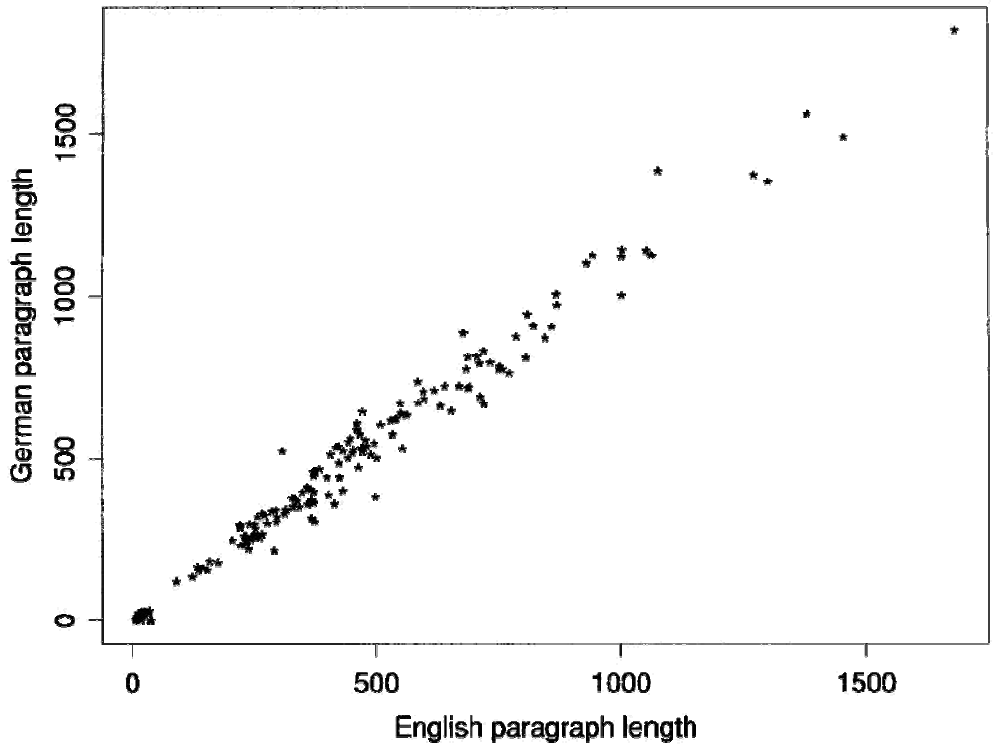


句子对齐的方法

- 句子对齐的基本方法
  - 基于长度的对齐方法
    - Brown等人的工作(1991)
    - Gale等人的工作(1993)
  - 基于单词的对齐方法
    - Kay等人的工作(1993)
- 两种方法对齐准确率都较高，对一般文本，都在90%以上。
- 基于长度的对齐方法效率优于基于单词的对齐方法。
- 基于单词的对齐方法：利用单词的对应关系，来决定句子的对齐关系。

基本依据

- 依据：
  - 互为翻译的两个句子在长度上高度相关。
  - 翻译时，句子顺序不做剧烈改变。（不考虑交叉）



什么是词语对齐？

- 在互为译文的两个句子间寻找词语对译关系。
- 形式定义：
  令 $S=s_1 s_2 \ldots s_J$ 代表原文句子
  令 $T=t_1 t_2 \ldots t_I$ 代表译文句子，
  则 二者间词汇级对齐A可定义为

$$A \subseteq \{s_1, s_2, \ldots, s_J\} \times \{t_1, t_2, \ldots t_I\}$$

  或者

$$A \subseteq \{(j,i) \mid j \in \{1,2,\ldots,J\}, i \in \{1,2,\ldots I\}\}$$

- 过于一般化

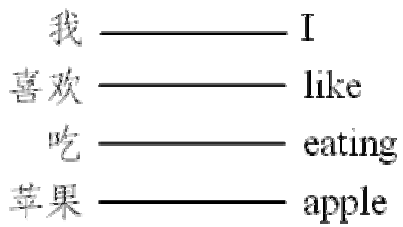
什么是词语对齐？

- 限制条件：不允许一对多的对应关系
- 原文中未译的词对应一个特殊的空词t_0
- 词汇对齐A是从集合$\{1,2, \ldots, J\}$到$\{0, 1, 2, \ldots, I\}$的映射

令$a_j=A(j)$，则：

$A = a_1 a_2 \ldots a_J$

- 词语对齐举例

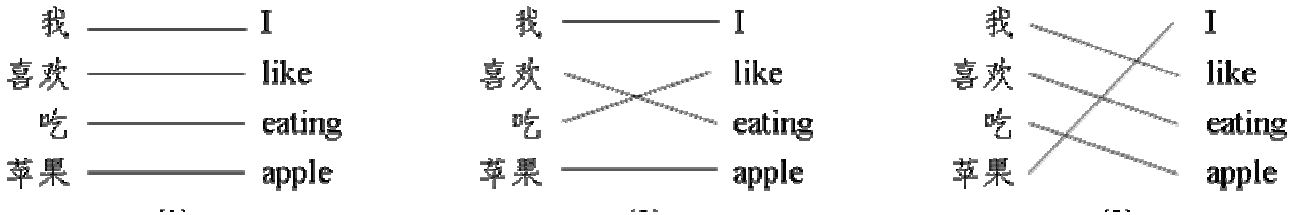


词语对齐

- 词语对齐较句子对齐困难
  - 翻译时，词序发生剧烈变化
  - 对应情况复杂
  - 对应关系难以确定（虚词）
- 词语对齐的基本方法
  - 统计模型法
    - 建立统计对齐的数学模型
  - 启发式方法
    - 不一定建立对齐模型，运用假设-检验等技术

词语对齐

- 从统计角度看，所有的对齐都是可能的，只不过概率大小不同



- 原文句子、译文句子长度分别是J、I，共有多少可能的对齐？

求解韦特比对齐

- 可以通过下面的过程计算韦特比对齐
  1) 罗列出原文句子和译文句子间所有可能的对齐
  2) 对每一种对齐，计算$P(S, A|T)$
  3) 寻找能使$P(S, A|T)$ 取得最大值的A作为韦特比对齐

- 问题一：如何计算$P(S, A|T)$?
- 问题二：罗列所有对齐效率如何？

- EM算法 GIZA++

机器翻译评价

- 科学客观的评价往往是推动技术发展的重要因素。
- 机器翻译困难，机器翻译评价也不容易
- 广义的机器翻译评价
  - （1）翻译系统的译文质量，译文是否是可以理解的或可以出版的？
  - （2）翻译系统的效率，每分钟系统可以完成多少字的翻译？
  - （3）翻译系统的健壮性，系统是否可以健壮的处理任何文本，是否经常出现系统崩溃？
  - （4）系统界面的友好性，用户是否可以很容易的使用系统？

机器翻译评价

- 狭义而言，机器翻译的评价一般仅指机器译文质量的评价或译文质量的自动评价
- 评价标准
  - “信、达、雅”不能作为标准
  - 最常用的两个标准源于ALPAC报告
    - 译文的可理解性(Intelligibility)
      译文可在多大程度上为不懂原文的人所理解
    - 译文的忠实度(Fidelity)
      译文和原文在内容上有多大差异。
  - 可理解性、忠实度原则上相互独立，但事实上经常相关

机器翻译评价

- 日本的长尾真教授在评测日本科学技术厅机器翻译项目(Mu)的日英系统译文质量时，为可理解性和忠实度进行了分级。
- 可理解性：
  - (1) 译文意义明确。用词、语法、文体都贴切，无需修改。
  - (2) 译文可以理解。用词、语法、文体方面多少有些问题。不过这些缺点很容易由人修正。
  - (3) 译文的意义大体上可以把握，但对一些细节有疑问，需问问懂原文的人。
  - (4) 译文质量差，用词、语法问题多。下功夫思考之后，可以猜想原文的大致意思。修改这样的译文还不如重译。
  - (5) 译文完全不能理解。

机器翻译评价

- 忠实度
  - (1) 译文忠实的传达了原文的全部内容。
  - (2) 译文忠实的传达了原文的全部内容，只需少量修正。
  - (3) 译文基本上忠实的译出了原文内容，但需进行像调整词序那样的修改。
  - (4) 译文基本上忠实的传达了原文内容，但是短语间的关系、时态、单复数、副词位置及其他方面有错误，因此译后编辑对句子结构要有些调整。
  - (5) 原文有一部分内容没译出来，或者短语、句子的搭配有错误。
  - (6) 原文的内容、结构没能较好的译出来，短语、子句有丢失。但译文大体上还算是一个句子。
  - (7) 译文不成句子，完全不能反映原文的内容和结构。

机器翻译的评价

- 人工评价
- 自动评价
  - 基于测试点的自动评价
  - 基于n-gram的自动评价(BLEU)

机器翻译研究的发展历程

- 1949年，Warren Weaver提倡MT研究
  - 翻译的过程可用解密过程(decoding)来类比
- 1954年，演示Georgetown系统
- 50年代末，Bar-Hillel 对MT研究的批评
  - 难以跨越的 “语义障碍”(semantic barrier)
- 60年代，ALPAC报告，MT研究转入低谷
  - 可理解性(Intelligibility)
  - 忠实度(Fidelity)

机器翻译研究的发展历程

- 70年代，机器翻译研究开始复苏
  - TAUM-METEO系统获得成功
  - 欧共体启用SYSTRAN系统
  - 人工智能、知识工程进展的影响。
- 80年代，机器翻译研究呈繁荣局面
  - 日本实施五国合作的ODA计划
  - 欧盟实施Eurotra计划
  - 多个公司推出了MT产品
  - 机器翻译方法的进展
- 90年代，机器翻译方法的多样化