



机器翻译

刘秉权

智能技术与自然语言处理研究室

新技术楼612室

Email: liubq@hit.edu.cn



主要内容

- 机器翻译简介
- 统计对齐
- 统计机器翻译
- 机器翻译中汉语处理的特殊问题



机器翻译(Machine Translation)

- 自动将文本或谈话内容从一种语言翻译为另一种语言，为NLP最重要的应用领域之一



《红楼梦》 片断翻译

- 源文：黛玉自在床上感念宝钗……，又听见窗外竹梢蕉叶之上，雨声淅沥，清寒透幕，不觉又滴下泪来。
- 译文：As she lay there alone, Dai-yu's thoughts turned to Bao-chai, Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry.



文学翻译涉及哪些问题？

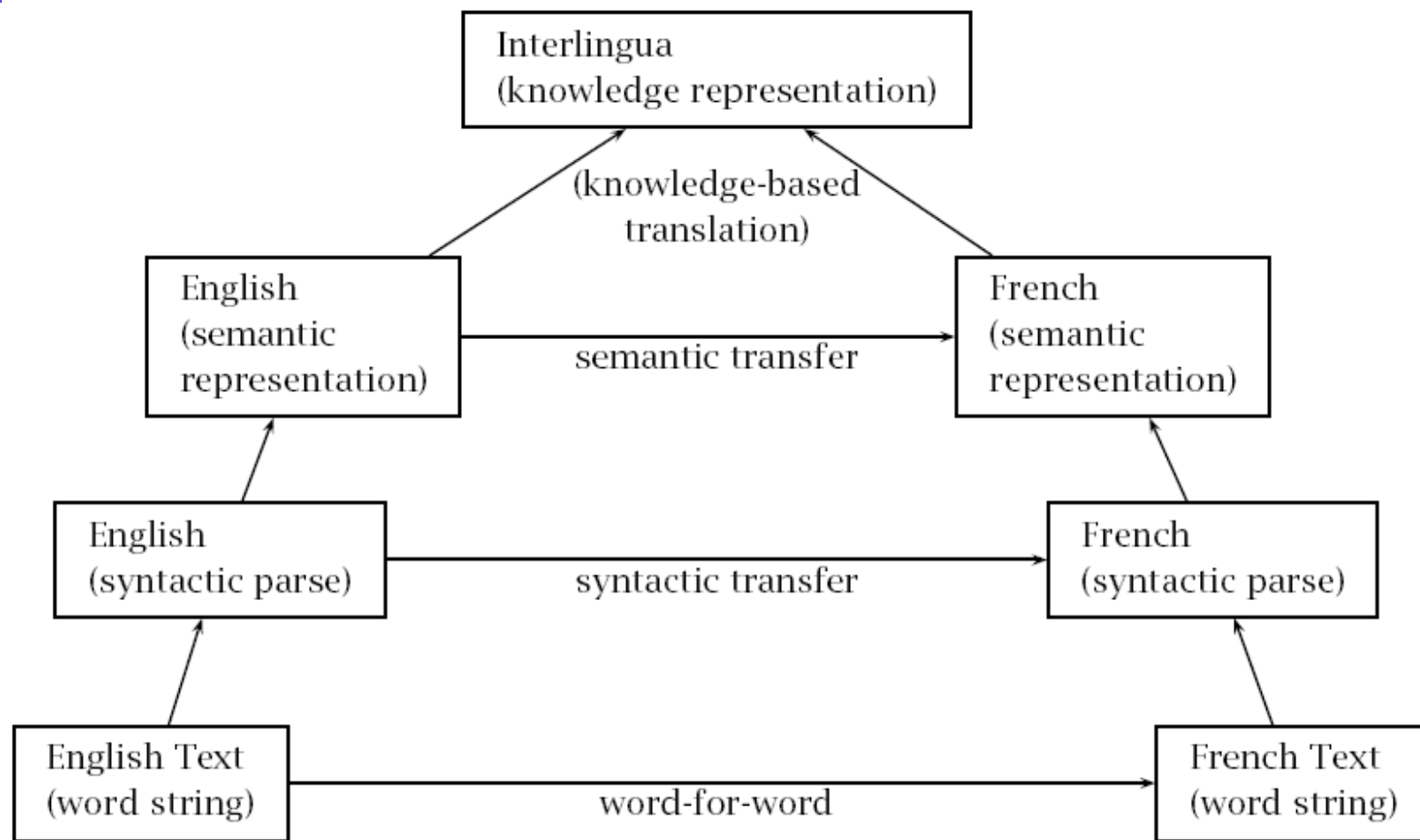
- 中文人名的翻译
 - 主要人名音译
 - 其他人意译：Aroma(袭人), Skybright(晴雯)
- 中文没有动词时态和语态变化
 - 透→penetrated
 - 幕→curtains of her bed
- 其他
 - 竹梢焦叶→bamboos and plantains



MT的难度和要求

- 高质量的翻译问题难以实现：对源语言和输入文本具有博大精深的理解，能够老练地、富有诗意地、创造性地支配目标语言
- 当前的计算模型可以胜任一些较简单的任务
 - 粗略翻译就足够的任务
 - 互联网中的“信息采集”
 - 人工编辑后可用于提高MT输出的任务
 - 机助翻译
 - 能够产生高质量译文的受限子语言领域的任务
 - 天气预报、航空旅行查询、约会安排、设备维护手册

MT的不同策略





直接翻译法：词-词对齐翻译

- 从源语言的表层句子出发，将词或固定词组直接置换成目标语言的对应成分
- 问题：对MT过程的认识过于简单
 - 不同语言之间可能不存在一一对应的映射关系
 - 词的歧义
 - 语言中的次序



句法转换法

- 解决了词序问题，一定程度上确保了翻译结果的句法准确性
- 问题：句法的正确性不等于语义的正确性
 - 德文短句 “Ich esse gern(I like to eat)”直译到英语结果为 “I eat readily”,英语中没有类似的 “动词-副词” 结构能表达 “I like to eat” 的概念，所以句法转换不能解决翻译中的所有问题



语义转换法

- 将原文转化为语义表示形式，在此基础上生成译文
- 能解决句法结构不匹配问题
- 问题：即使字面意思翻译完全准确，但最后译文对用户来说可能还是不易理解的



中间语言法

- 中间语言：独立于任何语言的知识表达形式
- 优点：进行多语种翻译时，只需对每种语言分别开发一个分析模块和一个生成模块
- 缺点：
 - 中间语言设计难度大：每种语言转化为中间语言都存在歧义
 - 语言本身的完备性构建也很困难



MT的主要方法

- 基于统计的方法(本章主要讨论内容)
- 统计与规则相结合的方法
- 基于实例的方法



统计对齐

- 文本对齐
- 词对齐



文本对齐

- 统计机器翻译的基础
- 相同的文字内容存在不同的语言版本
- 平行语料库(**parallel corpus**)
 - 官方文件：某些国家或地区具有多种官方语言(加拿大、瑞士、香港)
 - 数量大
 - 准确性高
 - 文学作品、宗教书籍
- 文本对齐：确定原文和译文句子或段落间的对应关系



句子和段落对齐

- 应用
 - 建立双语字典
 - 机器翻译
 - 多语言语料库的使用
 - 语义消歧
 - 多语言信息检索
- 对象：语言风格迥异、意译法等造成的不对齐现象(下页例子)

<i>With regard to</i>	Quant aux (à)	According to
<i>(the) mineral waters</i>	[(les) eaux	
<i>and (the) lemonades</i>	[minérales et aux	[our survey,] 1988
<i>(soft drinks)</i>	[(les) limonades,	
<i>they encounter</i>	[elles rencontrent	
<i>still more</i>	[toujours plus	[sales] of
<i>users</i>	[d'adeptes.	
<i>Indeed</i>	En effet	
<i>our survey</i>	[notre sondage]	[mineral water
<i>makes stand out</i>	fait ressortir	[and soft drinks] were
<i>the sales</i>	[des ventes]	
<i>clearly superior</i>	[nettement	[much higher]
	[supérieures]	[than in 1987,]
<i>to those in 1987</i>	[à celles de 1987,]	reflecting
<i>for cola-based</i>		[the growing popularity]
<i>drinks</i>	pour [les boissons à	of these products.
	[base de cola]	[Cola drink]
<i>especially</i>	[notamment.]	manufacturers
		[in particular]
		achieved above
		average growth rates.

对齐文本：
中间和右边两列分别是法文和英文句子，箭头标明了他们之间的对应关系，左边斜体字部分是从法文直译得到的英文翻译



句子对齐

- 定义：从句子内容出发，将源语言中的一组句子和目标语言中的一组句子对应的过程
- 每组句子可以为空，也可人为加入对应源语言中不存在的句子，或删除原有的句子
- 两组对应的句子为一个句珠(**bead**)
- 对齐方式：**1:1(90%)**、**1:n**、**n:1**、**m:n**
- 每个句子能且只能出现在一个句珠中
- 处理交叉依赖(**cross dependency**)问题



基于长度的对齐

- 基本原理：假设源语言和目标语言的句子长度存在比例关系
- 句子长度：定义为句子中单词或字符的个数
- 特点：简单、忽略很多其他可利用信息；效果好、效率高



统计对齐的目标

- 求概率最大的对齐

$$\arg \max_A P(A|S, T) = \arg \max_A P(A, S, T)$$

- 将对齐文本分解为句珠序列，各句珠之间独立分布

$$P(A, S, T) \approx \prod_{k=1}^K P(B_k)$$

- 在句珠内的句子已知的情况下，估算某一类句珠的概率值



一种基于长度的对齐算法

源语言: $S = (s_1, s_2, \dots, s_I)$, 目标语言: $T = (t_1, t_2, \dots, t_J)$

对齐方式: A , 句珠序列: (B_1, B_2, \dots, B_K)

i, j 代表两组句子 s_1, s_2, \dots, s_i 和 t_1, t_2, \dots, t_j

最小耗费函数: $D = (i, j)$

句子长度: l_1, l_2 , 两组句子的距离量度: δ

两组句子之间的耗费计算函数: $\text{cost}()$



实例：

利用句子长度评估两组句子的对齐程度

- 句子长度用其字符数表示
- 前提：段落对齐
- 假设只存在以下几类句珠模式：
 $\{1:1, 1:0, 0:1, 2:1, 1:2, 2:2\}$
- 利用动态规划寻找双语文本之间的“最短距离”
- 利用递归法反复计算最小耗费函数，找到其中的具备最小耗费的 $D(I, J)$

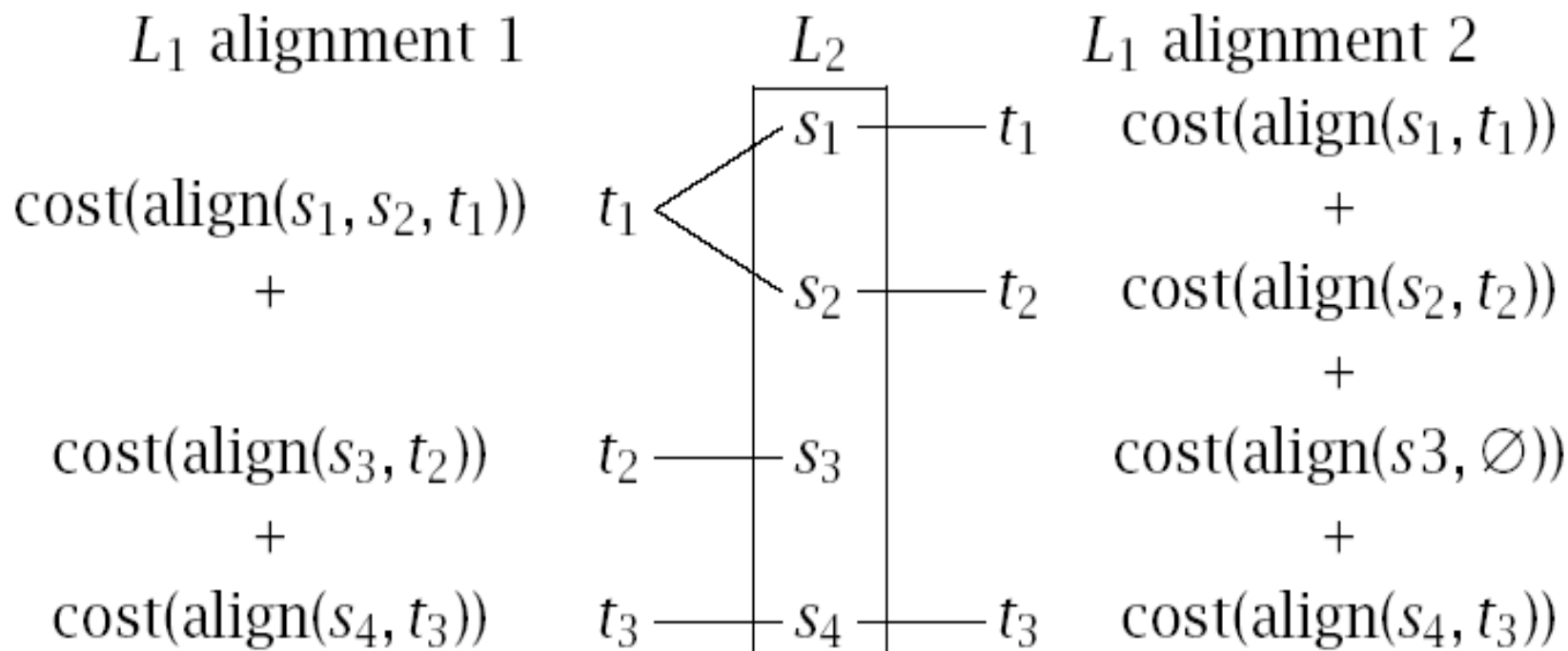


最小耗费函数计算

$$D(0, 0) = 0,$$

$$D(i, j) = \min \left\{ \begin{array}{l} D(i, j-1) + \text{cost}(0:1 \text{ align } \emptyset, t_j) \\ D(i-1, j) + \text{cost}(1:0 \text{ align } s_i, \emptyset) \\ D(i-1, j-1) + \text{cost}(1:1 \text{ align } s_i, t_j) \\ D(i-1, j-2) + \text{cost}(1:2 \text{ align } s_i, t_{j-1}, t_j) \\ D(i-2, j-1) + \text{cost}(2:1 \text{ align } s_{i-1}, s_i, t_j) \\ D(i-2, j-2) + \text{cost}(2:2 \text{ align } s_{i-1}, s_i, t_{j-1}, t_j) \end{array} \right.$$

计算对齐耗费的例子





确定每类对齐的耗费

耗费的计算基于句珠中每一种语言句子的长度 l_1 和 l_2 ，同时把一种语言中的每个字符相对于另一种语言对应出现的字符数作为随机变量，假设其服从正态分布，且满足独立同分布。从实际语料中估计这些参数，以均值 μ 为例，其值大概为 1，其方差 s^2 用长度差值的平方近似。



耗费计算函数

- 一个句珠内两组句子的距离度量

$$\delta = (l_2 - l_1\mu) / \sqrt{l_1 s^2}$$

- 耗费计算函数

$$\text{cost}(l_1, l_2) = -\log P(\alpha \text{ align} | \delta(l_1, l_2, \mu, s^2))$$

其中 $\alpha \text{ align}$ 代表某种可能的对齐模式，取负使得出的概率值和距离成反比；上式可通过贝叶斯公式计算，即 $P(\alpha \text{ align})P(\delta | \alpha \text{ align})$ ， $P(\alpha \text{ align})$ 相当于对齐模式的先验概率。



实验结果

- 此算法尽可能地在长度相近的句子之间建立了对齐关系
- 在UBS (Union Bank of Switzerland)语料库上进行实验
- 一般错误率4%
- 改进后达到0.7%
- 在“1:1”对齐模型上性能最好



其他方法

- 基于信号处理技术的偏移位置对齐算法
 - 在平行文本中利用位置偏移量的概念
 - 源文本中一定位置的文本和目标语言中一定位置的文本是大致对齐的
 - 可有效地处理噪声文本
- 句子对齐的词汇方法
 - 使用词语级别的部分对齐来推导出句子级别的对齐的最大似然估计方法，同时能反过来优化词语对齐
 - 假设：如果两个词语的分布相同，则它们是对应的
 - 不需要更高级别的段落对齐

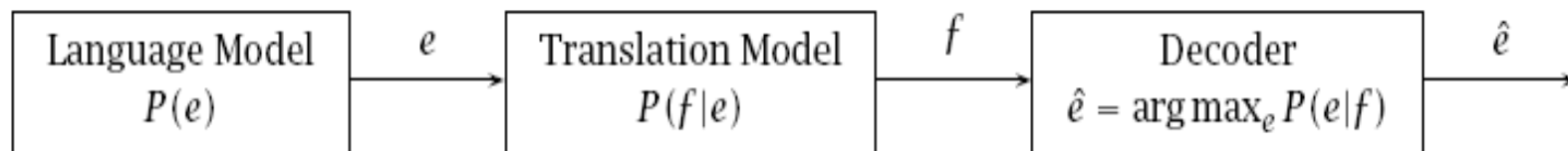


词对齐

- 利用词对出现的频率
- 利用关联度量的方法
- **EM**算法
- 更多地利用先验知识

统计机器翻译

■ 噪声信道模型(法文→英文翻译)



The noisy channel model in machine translation. The Language Model generates an English sentence e . The Translation Model transmits e as the French sentence f . The decoder finds the English sentence \hat{e} which is most likely to have given rise to f .



三个部分

- 语言模型
- 翻译模型
- 解码器
- 任务：参数值估计



语言模型(Language Model)

- 语言模型给出了英文句子的生成概率 $P(e)$
- 假定事先已经有一个定义良好的语言模型(n 元语法模型, 概率句法分析模型等)



翻译模型(Translation Model)

- 一个简单的基于词对齐的翻译模型

$$P(f|e) = \frac{1}{Z} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m P(f_j | e_{a_j})$$

其中： e 代表英文句子， l 是按词计算的句子长度； f 是法文句子， m 是其长度； f_j 是 f 的第 j 个词， a_j 是 e 中与 f_j 对齐的词的位置； e_{a_j} 是 e 中与 f_j 对齐的词； $P(w_f | w_e)$ 是翻译概率，即在已知 w_e 在英文句子里的条件下， w_f 出现在对应法文句子里的概率， Z 是归一化常数。



基本思想

- 它累加了法语单词对齐英语单词所有情况的概率， a_j 等于0表示法语句子中的第j个单词在英文中没有对齐成分
- 每个英文单词可以对应于多个法语单词，但每个法语单词至多仅和一个英文单词对应
- 各个单词的翻译之间是无关的
- 句子对齐后，对所有m个对齐概率求乘积作为对齐模式的概率



例

- 计算下面两句子的对齐概率

$$P(\textit{Jean aime Marie} | \textit{John loves Mary})$$

- 只需将三个单独的概率值相乘

$$P(\textit{Jean} | \textit{John}) \times P(\textit{aime} | \textit{loves}) \times P(\textit{Marie} | \textit{Mary})$$



解码器(Decoder)

$$\hat{e} = \arg \max_e P(e|f) = \arg \max_e \frac{P(e)P(f|e)}{P(f)} = \arg \max_e P(e)P(f|e)$$

- 启发式搜索算法
- 栈搜索法：渐进地构建英文句子，算法过程中随时保持一个局部的翻译，每次在这个局部翻译的基础上增加少数单词和对齐模式。如果当前扩展正确的可能性很小，可以通过剪裁栈返回以前的状态。
- 特点：效率高，但不能保证找到全局的最优翻译



翻译概率(Translation Probabilities)

- 用**EM**算法估计翻译模型的参数(假设句子已对齐)
- 基本思路：解决信任度分配问题。如果某个词和目标语言中的某个特定词对齐的可能性很大，则完全可以绑定它们，从而避免“1:2”或“1:3”的对齐模式及过多的没有对齐的词



简化算法

1. 随机设定 $P(w_f | w_e)$ 的初始值
2. E 步骤: 设 w_e 已知, 计算 w_f 在法语句子中出现次数的期望值:

$$z_{w_f, w_e} = \sum_{(e, f) \text{ s.t. } w_e \in e, w_f \in f} P(w_f | w_e)$$

(简单起见, 假设一个词在句子中仅出现一次)

3. M 步骤: 重新估计翻译模型参数:

$$P(w_f | w_e) \frac{z_{w_f, w_e}}{\sum_v z_{v, w_e}}$$



实验结果

■ 解码器导致的错误

- a. **Source sentence.** Permettez que je donne un exemple à la chambre.
- b. **Correct translation.** Let me give the House one example.
- c. **Incorrect decoding.** Let me give an example in the House.

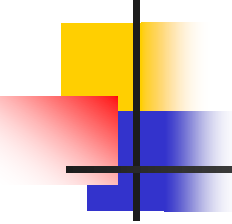
■ 语法结构解码错误

- a. **Source sentence.** Vous avez besoin de toute l'aide disponible.
- b. **Correct translation.** You need all the help you can get.
- c. **Ungrammatical decoding.** You need of the whole benefits available.



噪声信道模型面临的问题

- “富余度”的不对称性
 - 富余度：一种语言中一个单词对应另一种语言中单词的数目
- 独立性假设
- 对训练数据的敏感性
- 算法效率



另外一些问题

- 没有引入短语概念
- 非局部依存
- 词态变化。模型中具备词态变化的词被区分为不同的词，需单独学习
- 数据稀疏问题



讨论

- 问题根源所在：很少考虑自然语言领域本身的知识
- 统计翻译模型研究的未来重点：如何将语言中内在的语言规律性知识融入到模型中
- 其他非语言学模型在词对齐方面被证明也是有效的



机器翻译中汉语处理的特殊问题

- 汉语时态的处理。
- 汉语句子结构的分析。
- 特殊语言范畴的处理：无冠词对应结构的处理、被字句处理
- 汉语的零指代问题处理。
- 作为源语言和目标语言需考虑不同的处理。

谢谢！

