机器翻译

杨沐昀

教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

我们学了：子问题+模型+算法

- 绪论：数据（语言）的性质、粒度的产生 | 集合
- 语料库：数据的时空采样 和加工 | 分布统计、正太分布
- 汉语分词：词法分析、语言模型 | MM & HMM 动态规划
- 句法分析：CFG | PCFG | 移进-规约、chart、CYK
- 词义消歧：语义 | 共现、互信息、贝叶斯
- 篇章： 指代、衔接、连贯、RST、 | 理性主义 vs 经验主义{机器学习}
- Tasks：

目录

机器翻译概述

- 引言
- 机器翻译的产生与发展
- 机器翻译的主要研究内容及难点
- 小结

机器翻译概述

- 在线机器翻译系统

机器翻译概述

- 机器翻译的概念
  - 机器翻译 (machine translation, MT) 是用计算机把一种语言(源语言) 翻译成另一种语言(目标语言)的一门学科和技术

机器翻译概述

- 引言
  - 目前全世界总共存在5000余种语言;
  - 其中有总共有19种语言的使用人口达5000万;
  - 传统的人工翻译的方式早已无法满足人类文明发展的需求;
  - 机器翻译自计算机出现以来，一直是研究的热点;
    - 1947年，W. Weaver 就提出机器翻译备忘录

机器翻译的产生与发展

- 草创时期
  - 1947年，W. Weaver发表了以Translation为题目的备忘录，正式提出机器翻译问题。
  - Weaver 的两个基本观点：
    - 翻译类似于解读密码的过程；
    - 原文和译文“说的是同样的事情；
  - 第一个翻译系统的诞生
    - 1954年 Georgetown 大学在 IBM 协助下，实现了第一个俄译英MT系统；
    - 该系统只有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

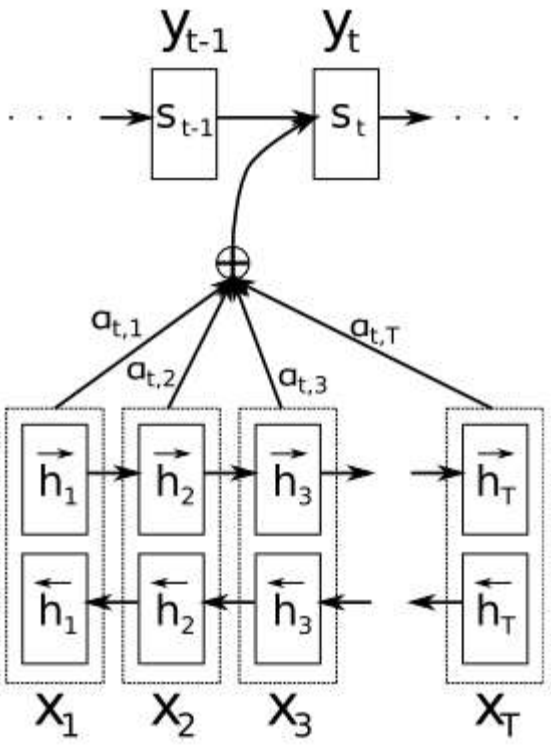
- 低谷时期：1970 ～1976年
  - ALPAC报告
    - 1964年，美国科学院成立语言自动处理咨询委员会 (ALPAC)，调查机器翻译的研究情况
    - 1966年 11月公布了一个题为“语言与机器”的报告，简称 ALPAC 报告;
    - 报告宣称：
      - “在目前给机器翻译以大力支持还没有多少理由”;
      - “机器翻译遇到了难以克服的语义障碍 (semantic barrier)”;
    - 从此，机器翻译研究在世界范围内进入低迷状态;

机器翻译的产生与发展

- 复苏与繁荣时期: (1976年～2013年)
  - 蒙特利尔大学与加拿大政府联合开发的机器翻译系统TAUM-METEO
    - 用于天气预报翻译。每小时可以翻译6～30万个词；
    - 每天翻译1500－2000 篇天气预报资料，并通过电视、报纸等立即公布；
  - 1978年欧共体启动多语言机器翻译计划
  - 1990年，IBM 提出基于词的统计机器翻译模型
    - 奠定机器翻译研究进入了繁荣期的基础。
  - 统计机器翻译模型的成熟期：基于短语的统计机器翻译模型
    - Tides计划/TIA
    - 开源统计机器翻译系统: Moses等

机器翻译的产生与发展

- 当前现状：神经机器翻译模型
  - 2014～至今
  - Sequence to sequence MT model；
  - Encoder-decoder based on attention;
  - Transformer结构成为目前主流的机器翻译模型



机器翻译的研究内容及难点

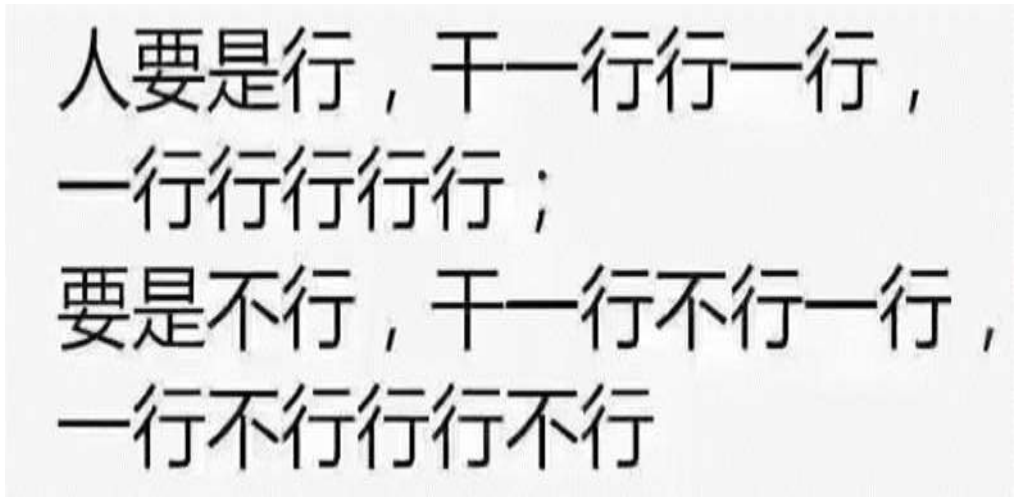
- 研究内容
  - 文本翻译：文本到文本的翻译;
  - 文本对齐：翻译过程中的词/短语对齐情况;
  - 质量评价：不给出参考译文的条件下评价机器译文的质量;
  - 译文后编辑：对机器译文进行人工后编辑；
  - ……

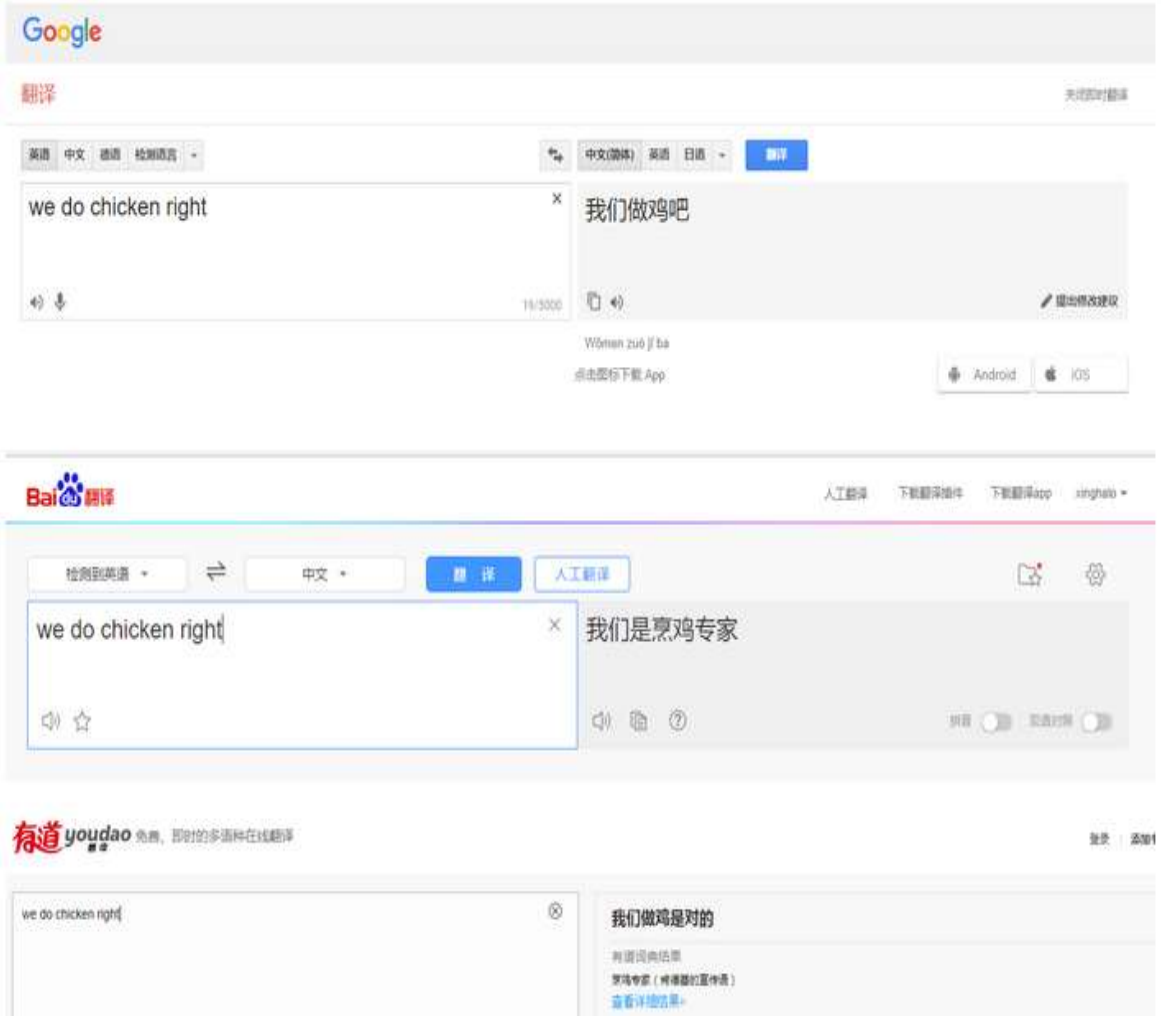
机器翻译的研究内容及难点

- 难点
  - 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 …
    - 新的词汇、术语、结构、语义 …
  - 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
  - 机器翻译的解不唯一且没有统一的标准

机器翻译的研究内容及难点

- 难以解决的问题：歧义
  - We do chicken right.
    - 我们做鸡肉是对的;
    - 我们马上做鸡肉;
    - 我们做右边的鸡肉;





机器翻译的研究内容及难点

- 难以解决的问题：新词/专有名词

  如食品或菜单名的翻译：

  馒头:

  Steamed bread ？

  Steamed bun

  夫妻肺片：

  Fuqifeipian/ Spicy beef

  童子鸡：

  Spring chicken/ Broiler chicken

  ……

机器翻译概述（小结）

- 协助而非替代
  - 目前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]
- 机器翻译还不成熟，需要的是人与系统的配合
  - 辅助机器翻译可以大大减轻人的负担；
- “信、达、雅”是翻译追求的目标
  - 计算机在这方面难以替代人；

传统机器翻译方法

- 直接转换法
- 基于规则的机器翻译方法
- 基于实例的机器翻译方法
- 其他翻译方法
- 小结

传统机器翻译方法

- 直接翻译法
  - 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。
  - 举例

I like Mary. → Me(I) gusta(like) Maria(Mary).

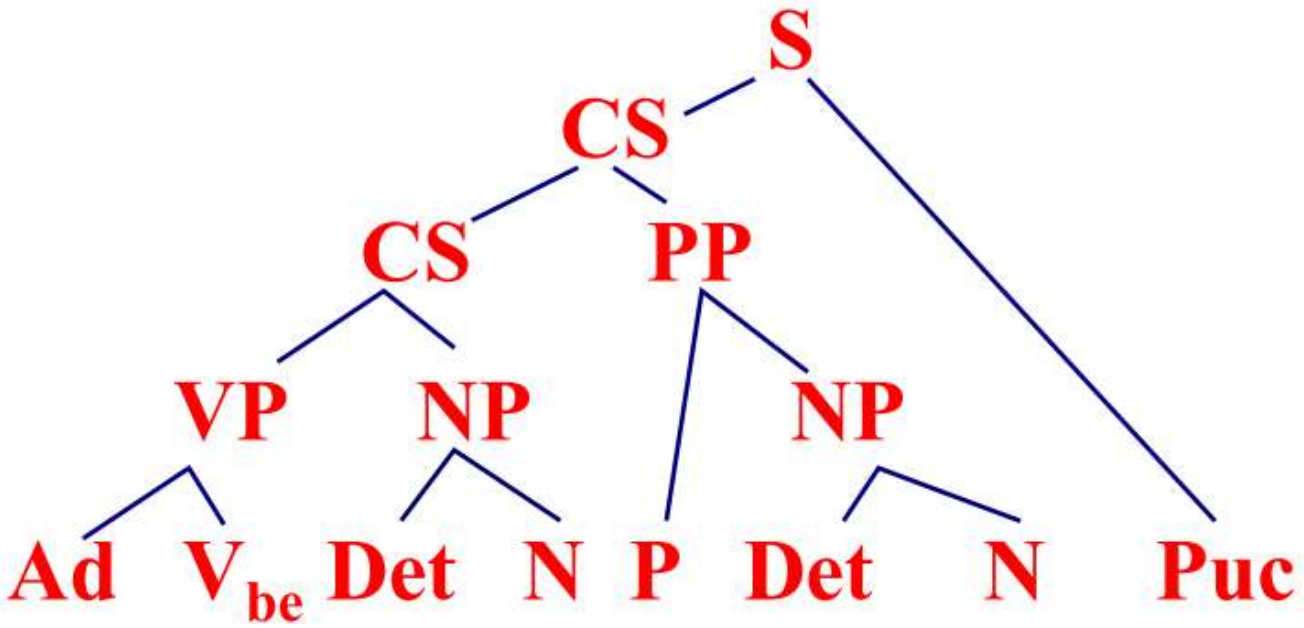
X like Y → Y X gusta

基于规则的翻译方法

- 定义：
  - 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
  - 用规则描述语法的翻译方法
- 基本过程
  - (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

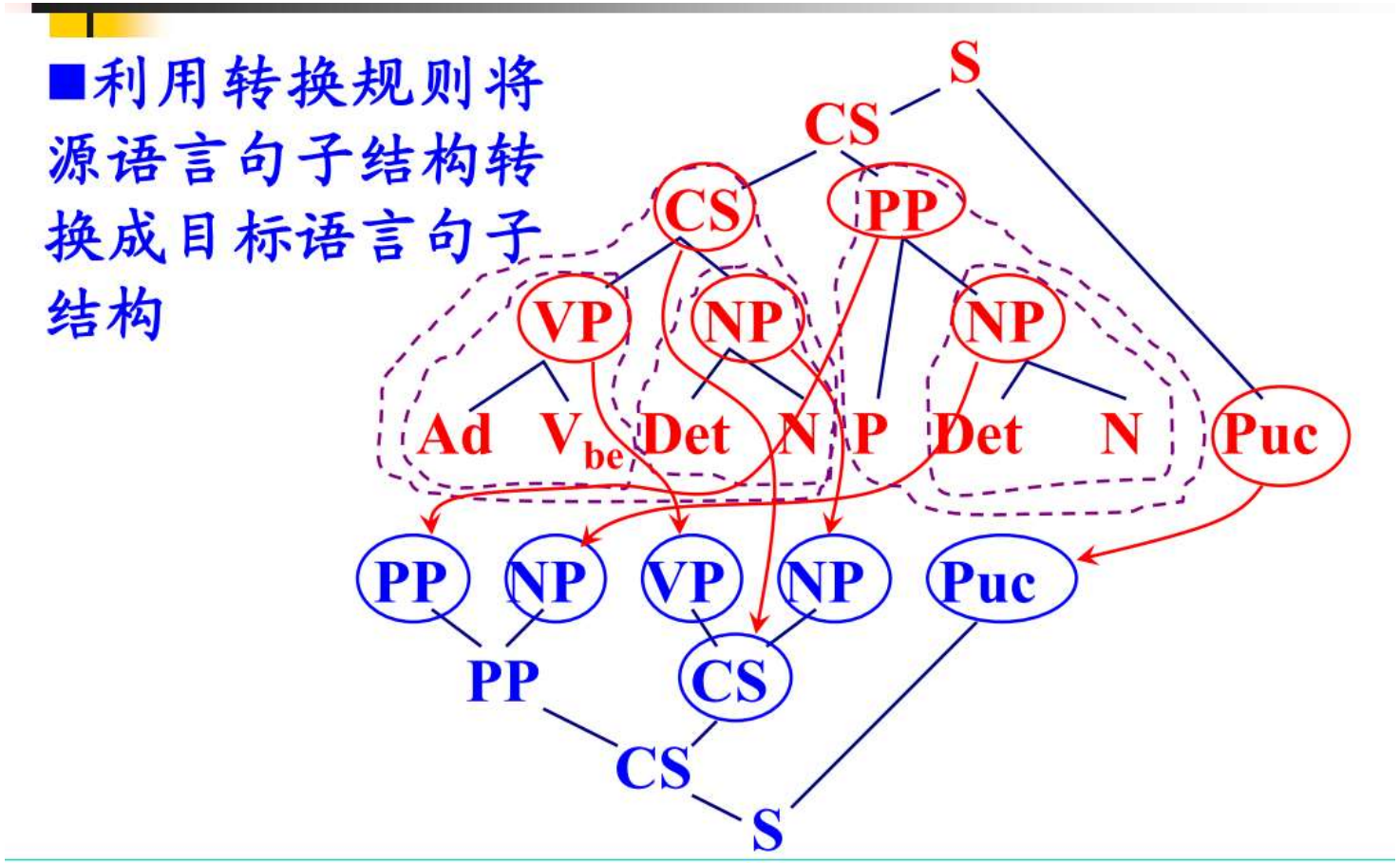
基于规则的翻译方法

- 举例
  - 给定源语言句子 There is a book on the desk.
  - 词法分析：There/Ad is/Vbe a/Det book/N on/P the/Det desk/N ./Puc
  - 利用句法规则进行句法结构分析：



基于规则的翻译方法

- 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



基于规则的翻译方法

- 将源语言词汇翻译成目标语言词汇
  - # there Ad: 在那里
  - # be Vbe: 是
  - # there be VP: 在…有
  - # a Det: 一，一个，一本…
  - # book N: 书，书籍; V: 预订
- 译文词法处理和目标语言句子生成：
  - 在桌子上有一本书。

基于规则的翻译方法

- 执行过程
  - “独立分析－独立生成－相关转换” 又称基于转换的翻译方法
- 代表系统
  - ARIANE翻译系统：
    - 由法国格勒诺布尔(Grenoble)机器翻译研究所(GETA)开发
  - TAU-METEO：天气预报信息服务
    - 1976年蒙特利尔大学与加拿大联邦翻译局联合开发的实用性机器翻译系统

基于规则的翻译方法

- 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
    - 尤其对于语言现象的或句法结构的明确的源语言语句具有较强的处理能力
  - 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。

基于实例的翻译方法

- 1984年由日本学者长尾真提出

EBMT is an acronym for
Example-Based Machine Translation.

Analogy-based,
Memory-based,
Pattern-based,
Case-based,
Similarity-based,
,
,
,
,

Prof. **Nagao Makoto**'s seminal paper
"Translation by analogy" **in 1981**.

Machine translation systems developed so far have a kind of inherent contradiction in themselves. **The more detailed a system has become by the additional improvements, the cleaner the limitation and the boundary will be made as for translation ability**. To break through this difficulty **we have to think about the mechanism of human translation**, and have to build a model based on the fundamental function of the language processing in human brain.

What is EBMT

"Translation by analogy."

(1) Man does **not** translate a simple sentence by doing deep **linguistic analysis**, rather,

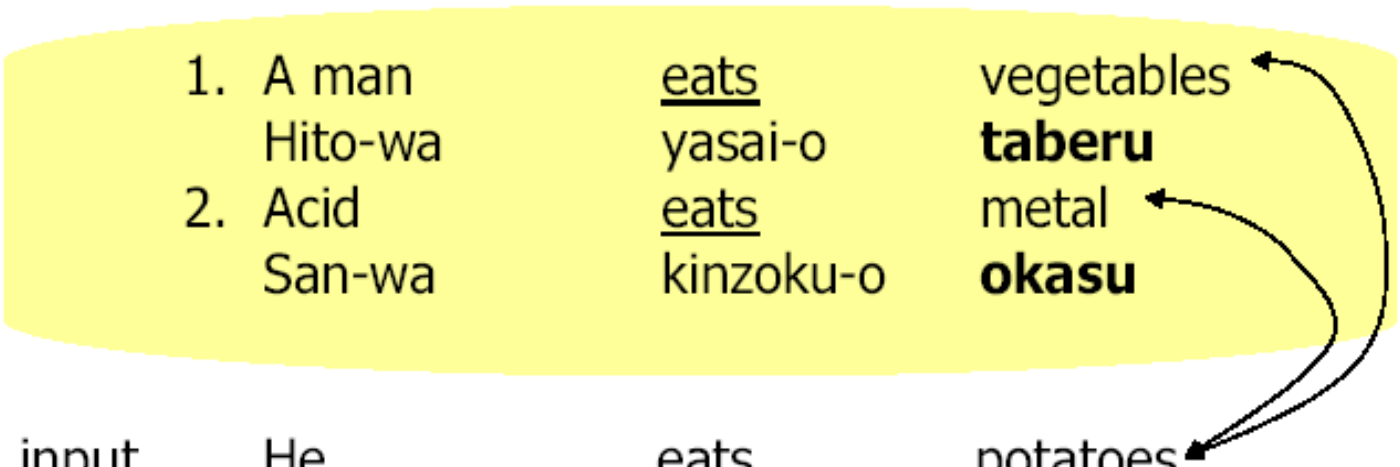
(2) Man does translation, first, by properly decomposing an input sentence into certain fragmental phrases…… The translation of each fragmental phrase will be done by the **analogy** translation principle **via proper examples** as its reference.

What is EBMT

- Nagao's Sample

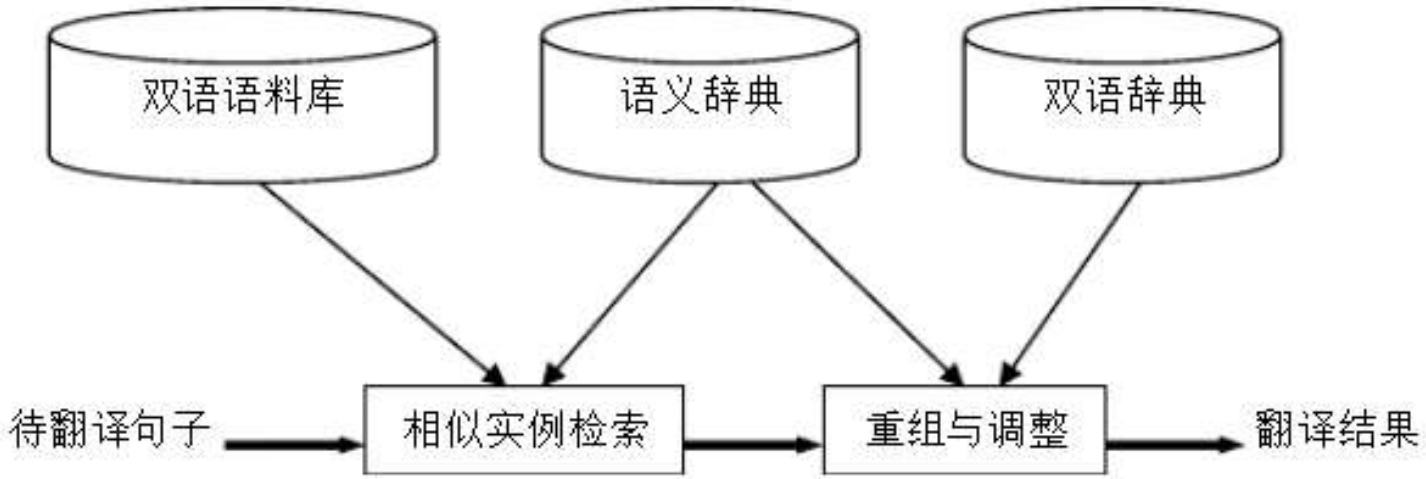
A selection of **Japanese** translations for the English word "eat"

1.	A man	eats	vegetables
	Hito-wa	yasai-o	**taberu**
2.	Acid	eats	metal
	San-wa	kinzoku-o	**okasu**
input	He	eats	potatoes
output	kare-wa	poteto-o	**taberu**



基于实例的翻译方法

- 方法概述
  - 输入语句 → 与事例相似度比较 → 翻译结果
  - 资源：大规模事例库
  - 代表系统：ATR-MATRIX (ATR, Japan)



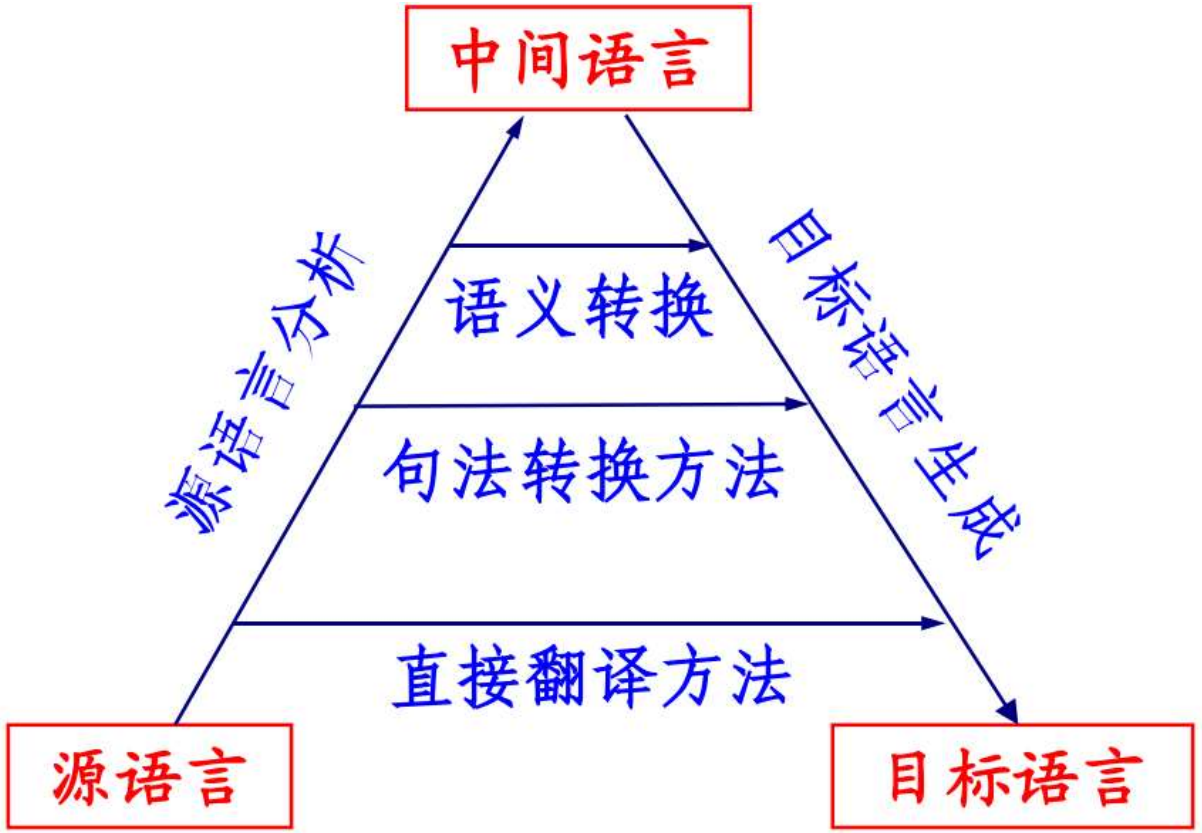
基于实例的翻译方法

- 举例
  - 待翻译句子：她 买了 一本 计算机语言学 入门书。
  - 实例库中已有翻译实例
    - 实例1：
    - 她买了 一件 时髦的 夹克衫。
    - She bought a sharp jacket.
    - 实例2：
    - 他 正在 读 一本 计算机语言学 入门书。
    - He has been reading a book on introduction to Computational Linguistics.
  - 翻译实例重组得到译文：
    - She bought a book on introduction to Computational Linguistics.

基于实例的翻译方法

- 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(包括结构相似性和语义相似性)往往难以把握
    - 在口语中，句子结构一般比较松散，成分冗余和成分省略都较严重;
    - 系统往往难以处理事例库中没有记录的陌生的语言现象；
    - 当事例库达到一定规模时，其事例检索的效率较低;

传统翻译方法小结



经典统计的机器翻译模型

- 统计机器翻译概述
- 基于词的机器翻译模型

统计机器翻译模型的发展历程

- 1947年 W. Weaver 提出“解读密码”的思想。
- 1990年IBM的Peter F. Brown 等在Computational Linguistics 上发表的论文“统计机器翻译方法”
- Brown等人发表的 “机器翻译的数学方法：参数估计”奠定了统计机器翻译的理论基础。
- Candide [Berger, 1994]在ARPA组织的机器翻译评测中表现良好。

统计机器翻译模型的基本原理

- 噪声信道模型
  - 一种语言T 由于经过一个噪声信道而发生变形从而在信道的另一端呈现为另一种语言 S (信道意义上的输出，翻译意义上的源语言)。
  - 翻译问题可定义为：
    - 如何根据观察到的 S，恢复最为可能的T 问题。
    - 这种观点认为，任何一种语言的任何一个句子都有可能是另 外一种语言中的某个句子的译文，只是可能有所出入 [Brown et. al, 1990]

统计机器翻译模型的基本原理

源语言句子：$S = s_1^m \equiv s_1 s_2 \cdots s_m$

目标语言句子：$T = t_1^l \equiv t_1 t_2 \cdots t_l$

贝叶斯公式：$p(T|S) = \dfrac{p(T) \times p(S|T)}{p(S)}$

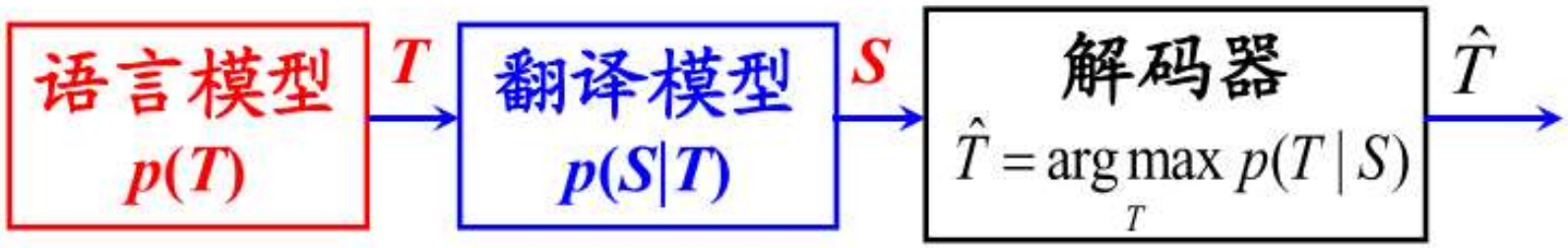
$$\hat{T} = \arg\max_T p(T) \times p(S|T)$$

语言模型
Language model, LM

翻译模型
Translation model, TM

统计机器翻译模型的基本原理

- 三个关键问题
  - (1)估计语言模型概率 p(T)；
  - (2)估计翻译概率 p(S|T)；
  - (3)快速有效地搜索T 使得 p(T)×p(S | T) 最大



统计机器翻译模型-语言模型

- 估计语言模型概率 p(T)

给定句子： $t_1^l = t_1 t_2 \cdots t_l$

句子概率： $p(t_1^l) = p(t_1) \times p(t_2 \mid t_1) \times \cdots \times p(t_l \mid t_1 t_2 \cdots t_{l-1})$

- 常见的语言模型：N-gram语言模型

统计机器翻译模型-翻译模型

- 翻译概率 p(S|T) 的计算
  - 定义目标语言句子与源语言句子的对应关系
  - 分类：根据刻画对应关系的语言粒度可分为
    - 基于词的翻译模型
    - 基于短语的翻译模型

基于词的统计机器翻译模型

- IBM模型1：词汇翻译（词对齐）
- IBM模型2：增加绝对对齐模型
- IBM模型3：引入繁衍率模型
- 课后阅读
  - IBM模型4：增加相对对齐模型
  - IBM模型5：修正缺陷

IBM 模型1-引入

- 基于词的翻译过程（人的翻译过程）
  - 查词典：寻找可能的翻译候选
    - 例如翻译词right，词典中right的翻译有：右边；立刻（+away）；权利；正义……
  - 在所有可能的翻译候选中找到最合适的翻译
    - 例如在” I can finish it right away” 中，根据上下文right应该翻译为立刻
  - 逐词翻译获得最终译文
    - ” I can finish it right away” 翻译为” 我能立刻完成它 ”

IBM 模型1-引入

- 基于词的翻译过程（机器翻译过程）
  - 收集语料：从平行语料中建立词典
  - 估计翻译概率: 极大似然估计
  - 词对齐: 建模对齐函数

IBM 模型1-定义

- 形式化定义
  - 目标语句子f={f_1, f_2, f_3, ……, f_{l_f},}, 句长为l_f
  - 源语言句子e={e_1, e_2, e_3, ……, e_{l_e},}, 句长为l_e
  - 对齐函数： 源语言词 e_j 到目标语词 f_i 的对齐函数a: j→l
- IBM 模型1
  - 参数ε是正则化常数

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

IBM 模型1-举例

- 翻译英语句子: the house is small 到法语句子

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$p(e,a|f) = \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein})$$
$$= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4$$
$$= 0.0028\epsilon$$

对齐是怎么来的呢？

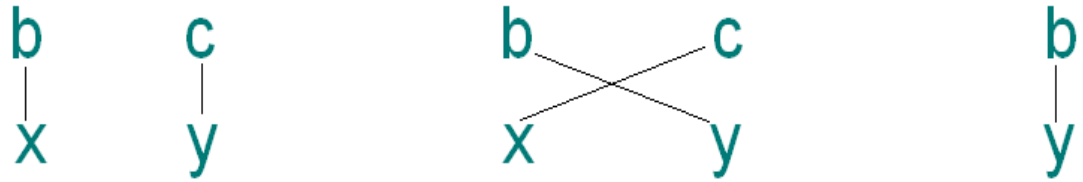
词对齐

- 从平行语料中估计词的翻译概率t(e|f)
  - 估计词的翻译概率需要词对齐信息
  - 得到词对齐信息需要知道词与词之间翻译概率
- Chicken-and-egg situation: 怎么解决？

EM 算法

EM 算法:Kevin Knight的例子

- Given two sentence beads:
  - b c ⇔x y
  - b ⇔y
- Possible alignments:

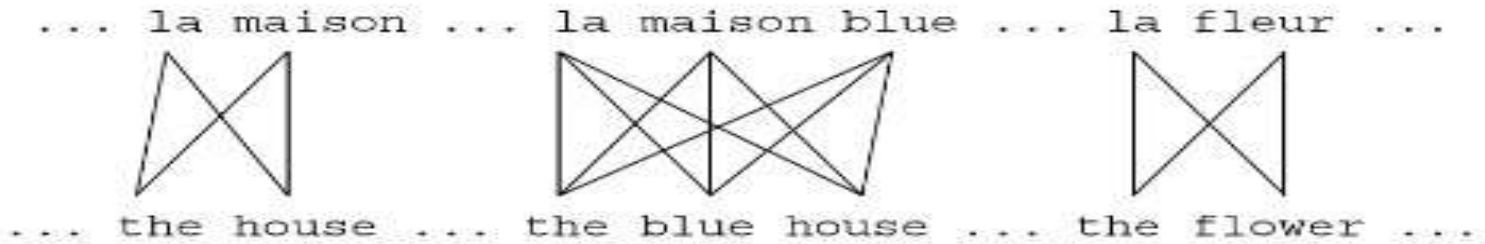


  - Magic:
    - Uniform distribute the translation probability:
      - p(x|b)=0.5 p(y|b)=0.5 | p(y|b)=1
    - Recount
      - p(x|b)= 0.5/ 2 ; p(y|b)=(0.5+1)/2
    - Repeat…… : p(x | b) = 0.0001 p(y | b) = 0.9999

EM 算法用于词对齐

- 初始步骤：所有词之间的对齐可能性相同
- 模型学习：将初始词对齐模型用于学习翻译概率

- EM算法(Expectation Maximization)



– 初始化时所有连接等效
– 统计发现，*la* 经常对齐到*the*

EM 算法用于词对齐

- 经过一轮迭代：根据得到的翻译概率更新对齐权重
- 对齐结果：la和the更有可能对齐

EM 算法用于词对齐

- 继续根据已有对齐结果进行迭代：更新对齐权重
- 对齐结果：出现更多的词对齐

EM 算法用于词对齐

- 直至收敛，得到所有词的对齐结果

EM 算法用于词对齐

- EM算法用于词对齐模型包含两步
- 期望步：将模型应用到数据
- 最大化步：利用数据来估计模型
- 直至收敛(无法继续更新)

(EM的有关理论证明，请参加机器学习等课程内容)

EM 算法用于词对齐-E步

- 计算对齐概率 $p(a|\mathbf{e},\mathbf{f})$
- 将问题转化为： $p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})}$
- 其中分子即为前述所定义的IBM模型1 $p(\mathbf{e},a|\mathbf{f}) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$
- 因此还需要计算分母 $p(\mathbf{e}|\mathbf{f})$

EM 算法用于词对齐-E步

- 计算

$$p(\mathbf{e}|\mathbf{f}) = \sum_a p(\mathbf{e}, a|\mathbf{f})$$

$$= \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} p(\mathbf{e}, a|\mathbf{f})$$

$$= \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)$$

EM 算法用于词对齐-E步

- 综上计算过程

$$p(\mathbf{a}|\mathbf{e},\mathbf{f}) = p(\mathbf{e},\mathbf{a}|\mathbf{f})/p(\mathbf{e}|\mathbf{f})$$

$$= \frac{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})}{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)}$$

$$= \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)}$$

EM 算法用于词对齐-M步

- 重新统计

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_{a} p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

- 代入前述结果

$$c(e|f;\mathbf{e},\mathbf{f}) = \frac{t(e|f)}{\sum_{i=0}^{l_f} t(e|f_i)} \sum_{j=1}^{l_e} \delta(e,e_j) \sum_{i=0}^{l_f} \delta(f,f_i)$$

- 估计模型

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}{\sum_{e}\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}$$

IBM算法1与词对齐-伪代码

Input: set of sentence pairs (**e**, **f**)
Output: translation prob. $t(e|f)$

initialize $t(e|f)$ uniformly
while not converged **do**
  // initialize
  count$(e|f) = 0$ **for all** e, f
  total$(f) = 0$ **for all** f
  for all sentence pairs (**e**,**f**) **do**
    // compute normalization
    for all words e in **e** **do**
      s-total$(e) = 0$
      for all words f in **f** **do**
        s-total(e) += $t(e|f)$
      end for
    end for
    // collect counts
    for all words e in **e** **do**
      for all words f in **f** **do**
        count$(e|f)$ += $\frac{t(e|f)}{\text{s-total}(e)}$
        total(f) += $\frac{t(e|f)}{\text{s-total}(e)}$
      end for
    end for
  end for
  // estimate probabilities
  for all foreign words f **do**
    for all English words e **do**
      $t(e|f) = \frac{\text{count}(e|f)}{\text{total}(f)}$
    end for
  end for
end while

IBM 模型1小结

- 基于词的统计翻译模型
  - 引入了词对齐的问题
  - 通过EM算法学习词对齐
- 缺陷
  - 无法刻画翻译过程中重排序、添词、舍词等情况；
  - 例如：
    - Seldom do I go to work by bus.
    - 我很少乘公共汽车上班。

IBM 模型2-引入绝对对齐

- 引入了对齐概率分布(alignment probabilities distribution)的概念

$$a(a_j \mid j,m,l) \equiv P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, l)$$

- 满足约束条件

$$\sum_{i=0}^{l} a(i \mid j,m,l) = 1$$

IBM 模型2

- 在IBM模型1的基础上进行优化

$$p(S|T)=\frac{\varepsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l}p(s_j|t_i)$$

引入绝对对齐

$$p(S|T)=\varepsilon\prod_{j=1}^{m}\sum_{i=0}^{l}p(s_j|t_i)\times a(i|j,m,l)$$

IBM 模型2-计算过程

- 根据 IBM模型2，由英语句子e 生成法语句子f 的实现过程：
  - 根据概率分布为法语句子f 选择一个长度m
  - 对于每一个 j = 1, 2, …, m，根据概率分布 $a(a_j \mid j,l,m)$ 从0, 1, …, l 中选择一个值给a_j
  - 对于每一个 j = 1, 2, …, m，根据概率选择一个法语单词f_j。

IBM 模型3-引入繁衍率

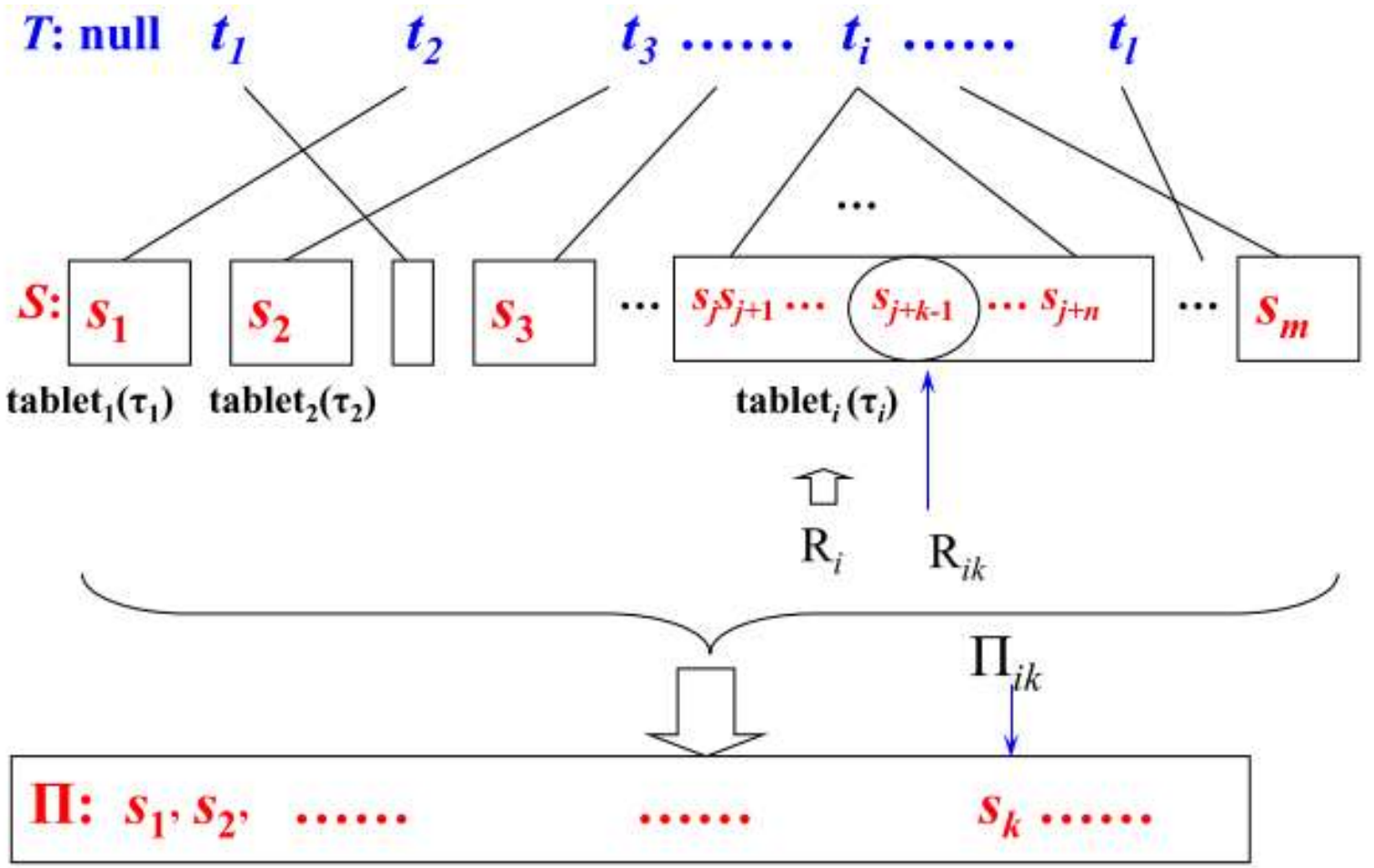
- 前述模型存在的问题
  - 在随机选择对位关系的情况下，与目标语言句子中的单词t对应的源语言句子中的单词数目是一个随机变量；
- 繁衍率：
  - 定义：与目标语言句子中的单词t对应的源语言句子中的单词数目的变量
  - 记做Φt，称该变量为单词t的繁衍能力或产出率(fertility)。一个具体的取值记做：Φt
  - 繁衍率刻画的是目标语言单词与源语言单词之间一对多的关系

IBM 模型3

- 片段(tablet)：
  - 假设给定一个目标语言句子T，T中的每一个单词t 在源语言句子中可能有若干个词与之对应，源语言句子中所有与t 对位的单词列表定义为t的一个片断；
- 片断集：
  - 这个片断可能为空。一个目标语言句子 T 的所有片断的集合是一个随机变量，我们称之为T 的片断集，记做符号R。
- T 的第i个单词的片段也是一个随机变量，不妨记做R_i，
- T 的第i个单词的片断中第k个源语言单词也是一个随机变量，记做R_{ik}。

IBM 模型3

- 模型示意图



IBM 模型3

标释集 τ(R的一个具体取值) 和单词排列 π(Π的一个具体取值，即τ中单词的一种排列方式) 的联合似然率为：

$$p(\tau,\pi\mid T)=\prod_{i=1}^{l}p(\varphi_i\mid\varphi_1^{i-1},T)\times p(\varphi_0\mid\varphi_1^{l},T)\times$$

繁衍概率 (fertility prob.)

$$\prod_{i=0}^{l}\prod_{k=1}^{\varphi_i}p(\tau_{ik}\mid\tau_{i1}^{k-1},\tau_0^{i-1},\varphi_0^{l},T)\times$$

翻译概率 (tran. prob.)

$$\prod_{i=1}^{l}\prod_{k=1}^{\varphi_i}p(\pi_{ik}\mid\pi_{i1}^{k-1},\pi_1^{i-1},\tau_0^{l},\varphi_0^{l},T)\times$$

位变概率 (distortion prob.)

$$\prod_{k=1}^{\varphi_0}p(\pi_{0k}\mid\pi_{01}^{k-1},\pi_1^{l},\tau_0^{l},\varphi_0^{l},T)\qquad(8)$$

IBM 模型3

假设：

(1) 对于1到l中的每一个i，概率 $p(\varphi_i \mid \varphi_1^{i-1}, T)$仅依赖于 ϕ_i 和 t_i，记作：$n(\varphi \mid t_i) \equiv p(\varphi \mid \varphi_1^{i-1}, T)$

(2) 对于所有的i，概率 $p(\tau_{ik} \mid \tau_{i1}^{k-1}, \tau_0^{i-1}, \varphi_0^l, T)$只依赖于$\tau_{ik}$和$t_i$，记作：$p(s \mid t_i) \equiv p(R_{ik} = s \mid \tau_{i1}^{k-1}, \tau_0^{k-1}, \varphi_0^l, T)$

(3) 对于1到l中的每一个i，$p(\pi_{ik} \mid \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \varphi_0^l, T)$只依赖于$\pi_{ik}$，$i, m$ 和l。位置概率记作：

$$d(j \mid i, m, l) \equiv p(\Pi_{ik} = j \mid \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \varphi_0^l, T)$$

IBM 模型3

设想τ_1^l中每个写字板中的一组词都存在一个额外词(即这个词在对齐时对空)，假设这个额外词出现的概率为p_1。另外，由于 $\phi_0+\phi_1+\cdots+\phi_l=m$，因此，

$$p(S|T)=\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}p(S,A|T)$$

$$=\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\binom{m-\phi_0}{\phi_0}(1-p_1)_0^{m-2\phi_0}p_1^{\phi_0}\prod_{i=1}^{l}\phi_i!n(\phi_i|t_i)$$

$$\times\prod_{j=1}^{m}p(s_j|t_{a_j})d(j|a_j,m,l) \qquad (9)$$

其中，

$$\sum_s p(s|t)=1$$

$$\sum_j d(j|i,m,l)=1$$

$$\sum_\phi n(\phi|t)=1$$

估计这些参数和 p_1。

IBM 模型3

- 计算过程

根据 IBM模型3, 一个英语句子e 翻译成法语句子f的工作过程如下：

(1) 对于英语句子中的每一个单词e，选择一个产出率ϕ，其概率为 $n(\phi|e)$;

(2) 对于所有单词的产出率求和，得到 *m-prime*;

(3) 按照下面的方式构造一个新的英语单词串：删除产出率为0的单词，复制产出率为1的单词，复制两遍产出率为2的单词，依此类推;

(4) 在这*m-prime*个单词的每一个后面，决定是否插入一个空单词NULL，插入的概率为p_1，不插入的概率为p_0;

IBM 模型3

- 计算过程

(5) 设ϕ_0为插入空单词 NULL 的个数；

(6) 设m为目前总单词的个数：$m\text{-}prime + \phi_0$；

(7) 根据概率表 $p(f|e)$ 将每一个单词e替换为法语单词f；

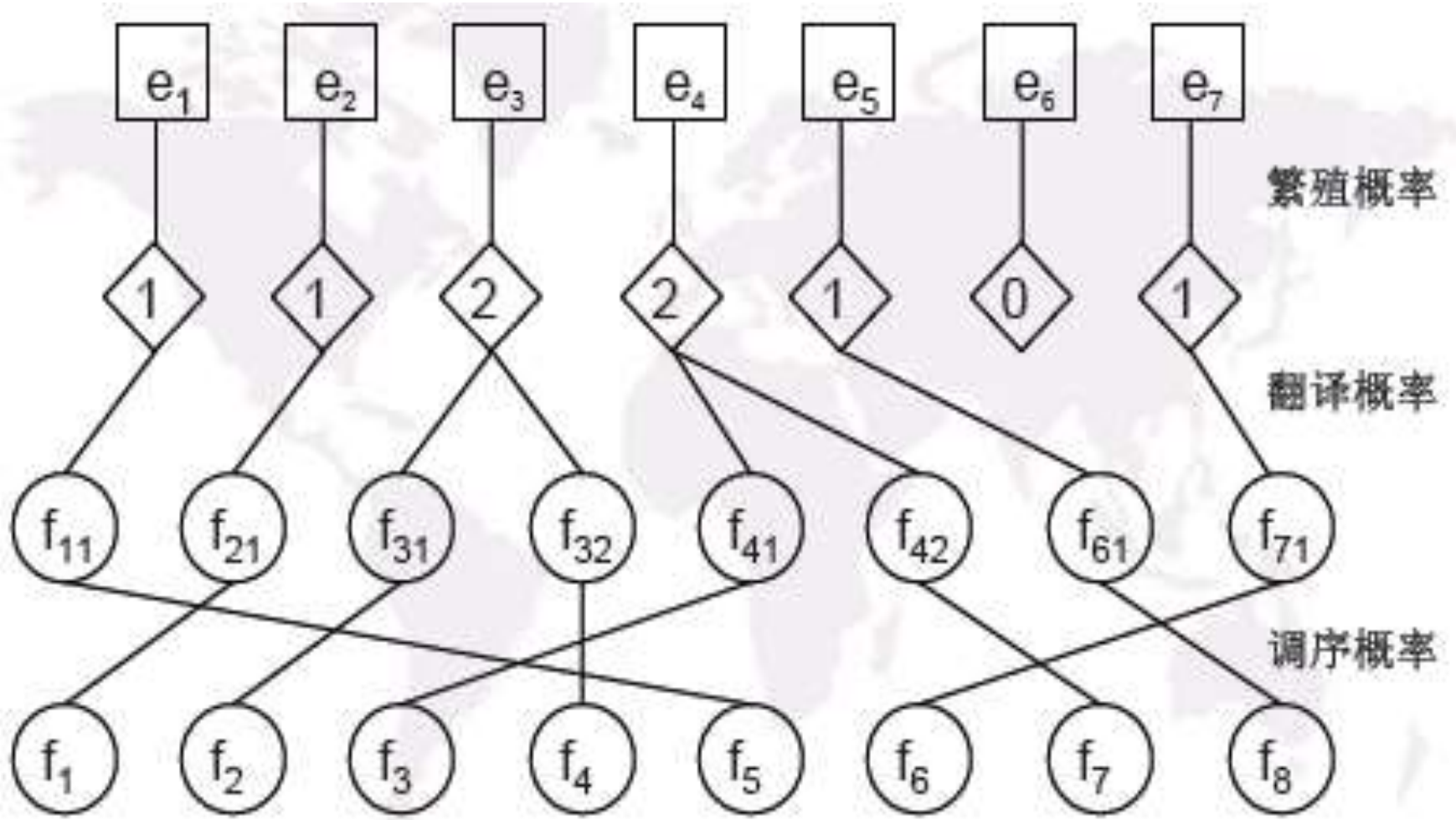
(8) 对不是由空单词 NULL 产生的每一个法语单词，根据概率表 $d(j\,|i, l, m)$ 赋予一个位置。这里j 是法语单词在法语句子中的位置，i是产生当前这个法语单词的英语单词在其句子中的位置，l是英语句子的长度，m 是法语句子的长度；

(9) 如果任何一个法语句子的位置被多重登录(含一个以上的单词)，则失败返回；

(10) 给空单词NULL产生的法语单词在句子中赋予一个位置，这些位置必须是没有被占领的空位置。任何一个赋值都被认为是等概率的，概率值为$1/\phi_0$；

(11) 读出法语单词串，其概率为上述每一步概率的乘积，按概率大小输出结果。

翻译模型:IBM Model 1-3



课后阅读-更高级的IBM

- IBM模型4：增加相对对齐模型
- IBM模型5：修正缺陷
- 参考书籍：

  《统计机器翻译》4.4节 Philipp Koehn著 宗成庆等译

IBM公司的Candide系统

- 基于统计的机器翻译方法
- •分析—转换—生成
  - –中间表示是线性的
  - –分析和生成都是可逆的
- •分析（预处理）：
  - 1.短语切分 2.专名与数词检测
  - 3.大小写与拼写校正
  - 4.形态分析 5.语言的归一化
- 转换（解码）：基于统计的机器翻译
- 解码分为两个阶段：
  - –第一阶段：使用粗糙模型的堆栈搜索
    - 输出140个评分最高的译文
    - 语言模型：三元语法
    - 翻译模型：EM Trained IBM Model 5
  - –第二阶段：使用精细模型的扰动搜索
    - 对第一阶段的输出结果先扩充，再重新评分
    - 语言模型：链语法
    - 翻译模型：最大熵翻译模型（选择译文词）

IBM公司的Candide系统

- ARPA的测试结果：

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

课后练习

- 阅读IBM模型2~5
  - 《统计机器翻译》4.4节 Philipp Koehn著 宗成庆等译
- 理解 IBM模型1的算法
- 运行使用开源词对齐工具GIZA++
  - 工具：http://www.fjoch.com/GIZA++.html
  - 语料: http://www.statmt.org/europarl/

延伸阅读列表

- Hutchins, W. J. and Somers, H. L. (1992). An Introduction to Machine Translation. Academic Press, London.
- Pierce, J. R. and Carroll, J. B. (1966). Languages and machines — computers in translation and linguistics. Technical report, Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences.
- Hutchins, W. J. (2007). Machine translation: a concise history. In Computer Aided Translation: Theory and Practice, C. S. Wai (ed.). Chinese University of Hong Kong.
- Brown, P. F., Cocke, J., Della-Pietra, S. A., et al. (1990). A statistical approach to machine translation. Computational Linguistics, 16(2):76–85.

延伸阅读列表

- Knight, K. (1999b). A statistical MT tutorial workbook. available at http://www.isi.edu/~knight/.
- Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation. Computational Linguistics, 19(2):263–313.
- Brown, P. F., Lai, J. C., and Mercer, R. L. (1991b). Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL).
- Wu, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL).

目录

基于短语的统计机器翻译模型

- 定义：最小翻译单元是“短语”
  - 此处的短语的概念不等同于语言学上的短语
  - 短语指一个连续的字串(n-gram)
- 引入：为什么要用基于短语的翻译模型？
  - 多对多翻译可以处理成非语言学的短语
  - 在翻译中可以使用局部上下文
  - 更多的数据：可以学习到较长的短语的翻译

基于短语的机器翻译模型

- 翻译过程举例

基于短语的机器翻译模型-模型训练

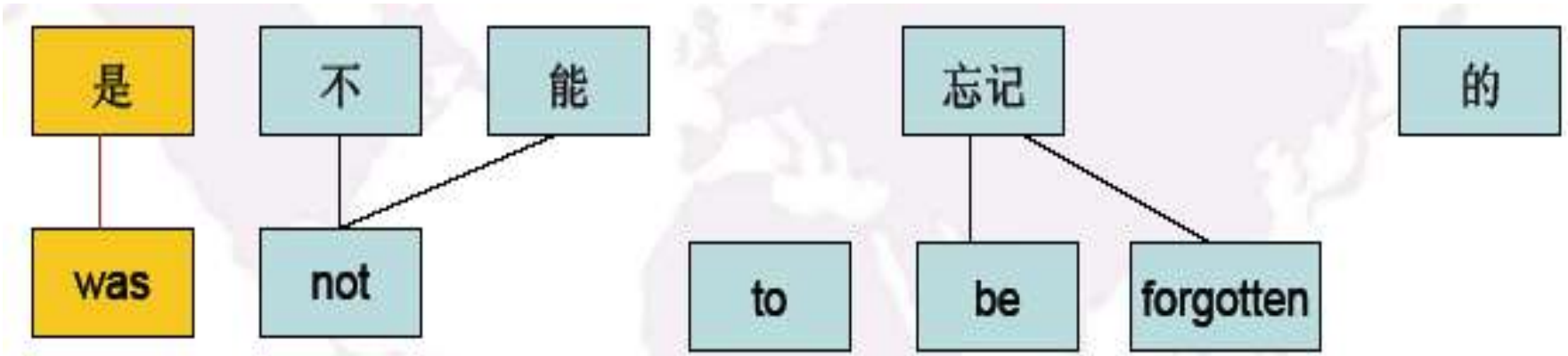
- 从平行语料中学习短语翻译表
  - 词对齐
  - 抽取短语对
  - 短语对打分

基于短语的翻译模型(1)

- 基本思想
  - 把训练语料库中所有对齐的短语及其翻译概率存储起来，作为一部带概率的短语词典
  - 这里所说的短语是任意连续的词串，不一定是一个独立的语言单位
  - 翻译的时候将输入的句子与短语词典进行匹配，选择最好的短语划分，将得到的短语译文重新排序，得到最优的译文.
- 问题：
  - 短语如何抽取？
  - 短语概率如何计算？

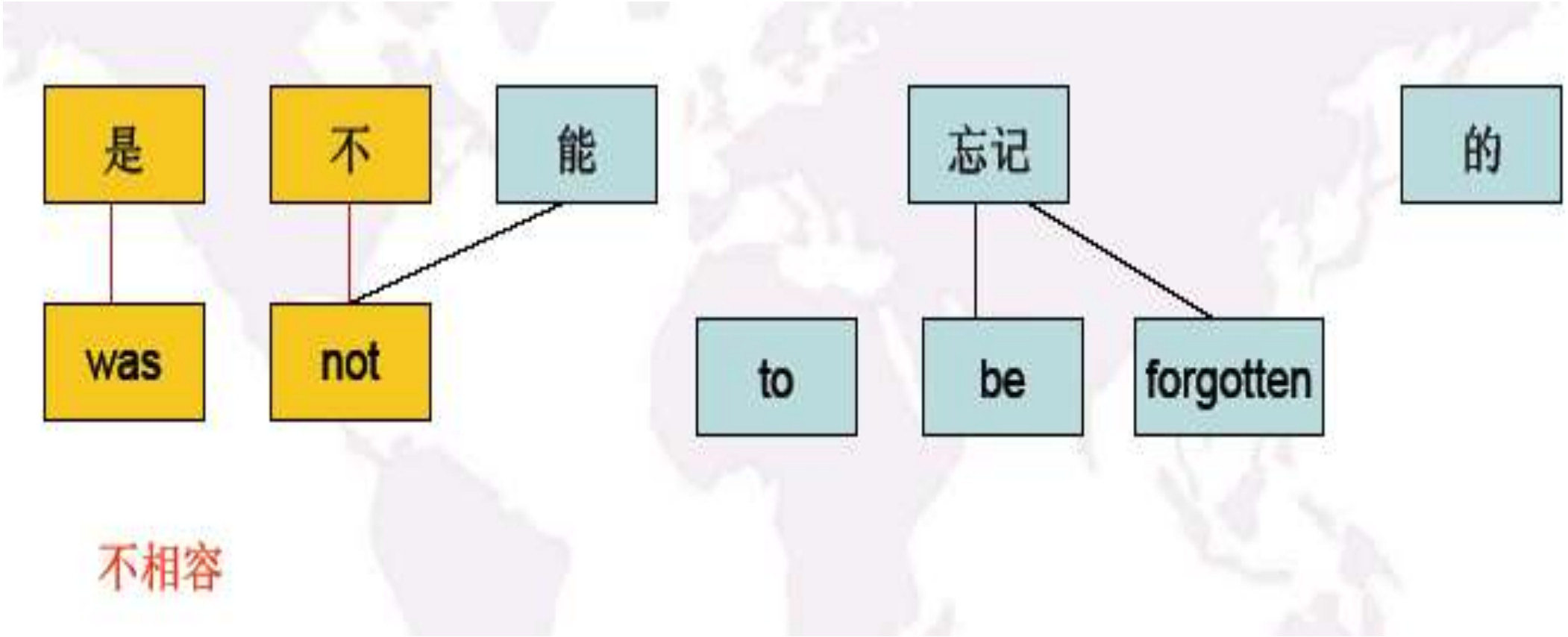
基于词语对齐的短语自动抽取(1)

- 列举源语言所有可能的短语，根据对齐检查相容性

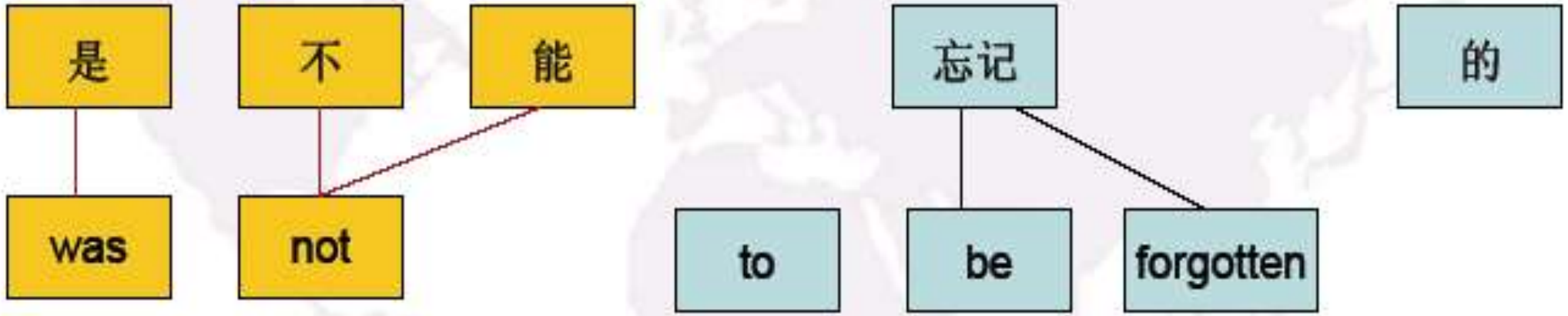


（是，was）

基于词语对齐的短语自动抽取(2)

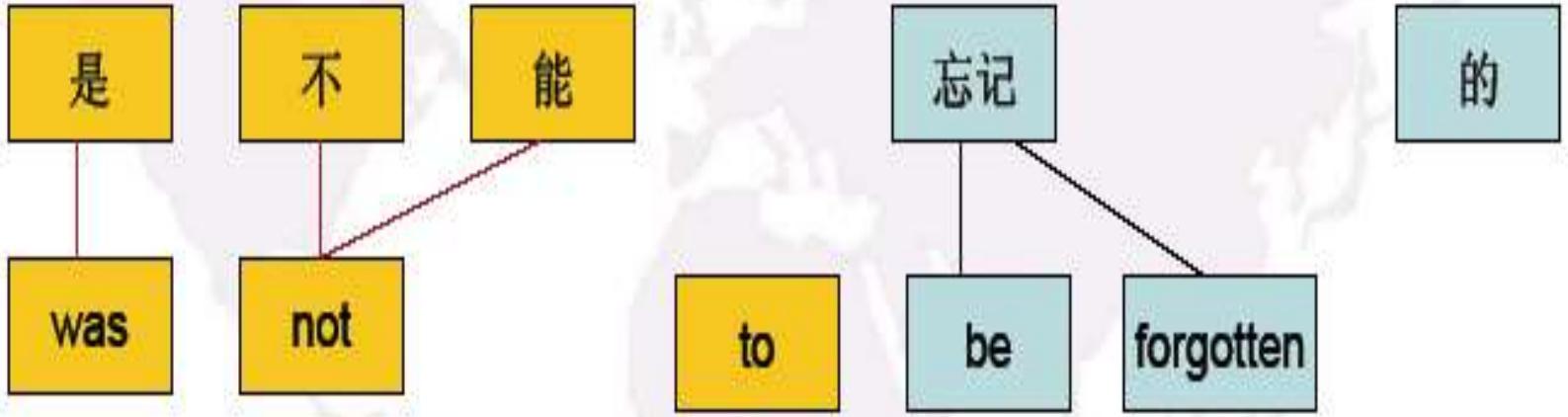


基于词语对齐的短语自动抽取(3)



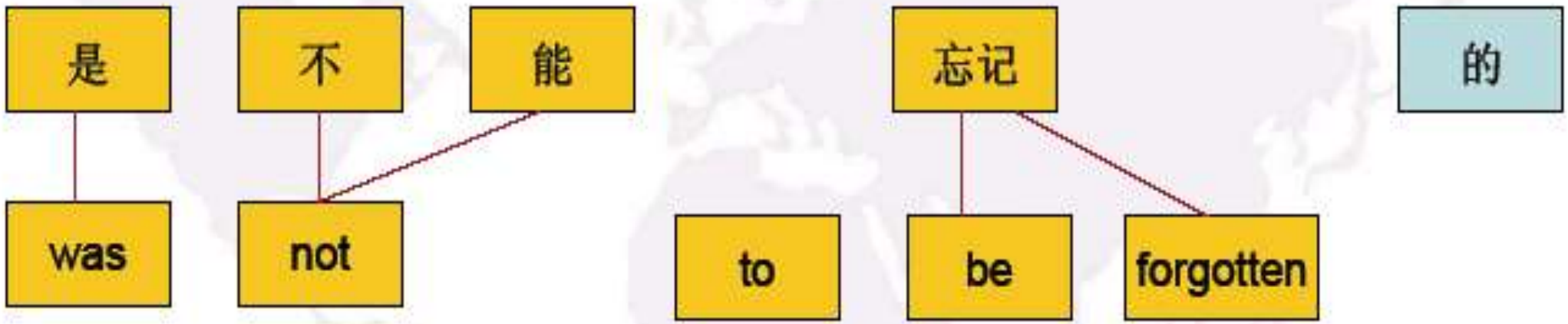
（是不能，was not)

基于词语对齐的短语自动抽取(4)



（是不能，was not to)

基于词语对齐的短语自动抽取(5)



（是不能忘记，was not to be forgotten)

基于词语对齐的短语自动抽取(6)

(是不能忘记的，was not to be forgotten)

基于词语对齐的短语自动抽取(7)

- 短语表

是	was
是不能	was not
是不能	was not to
是不能忘记	was not to be forgotten
是不能忘记的	was not to be forgotten
不 能	not
不 能	not to
不 能 忘记	not to be forgotten
不 能 忘记 的	not to be forgotten
忘记	be forgotten
忘记	to be forgotten
忘记 的	be forgotten
忘记 的	to be forgotten

基于短语的机器翻译模型-模型训练

- 短语对打分：为短语翻译分配概率
- 按相对频率打分：

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$

双语短语的概率计算(2)

- 前例：

不能 not
不能 not to
N(不能, not) = ½
N(不能, not to) = ½

如果语料库中另外有一句话：

N(不能, can not) = 1

则

$$p(not\ to \mid 不能) = \frac{1/2}{(1/2+1/2+1)} = 1/4$$

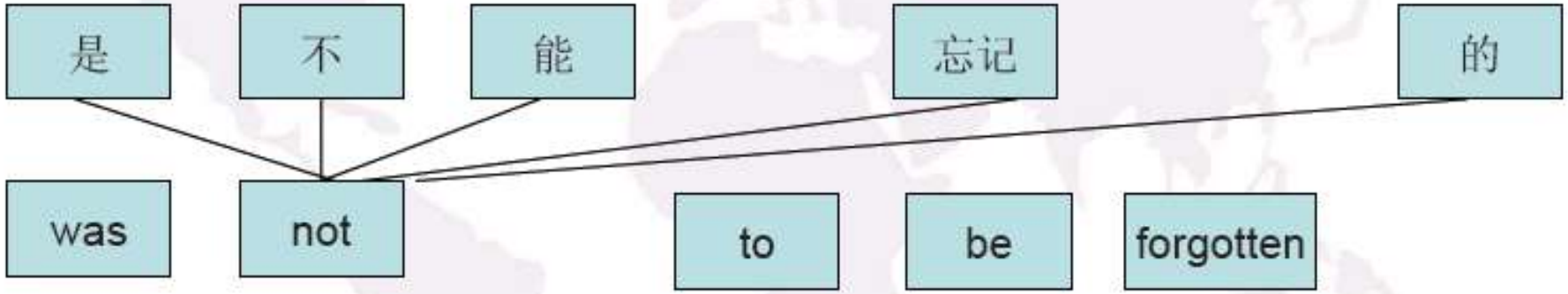
$$p(not \mid 不\ 能) = \frac{1/2}{(1/2+1/2+1)} = 1/4$$

$$p(can\ not \mid 不能) = \frac{1}{(1/2+1/2+1)} = 1/2$$

双语短语的概率计算(3)

- 仅利用共现次数计算概率信息不全面

  较短的短语（如单词）出现次数多，概率不集中，较长的短语概率比较大



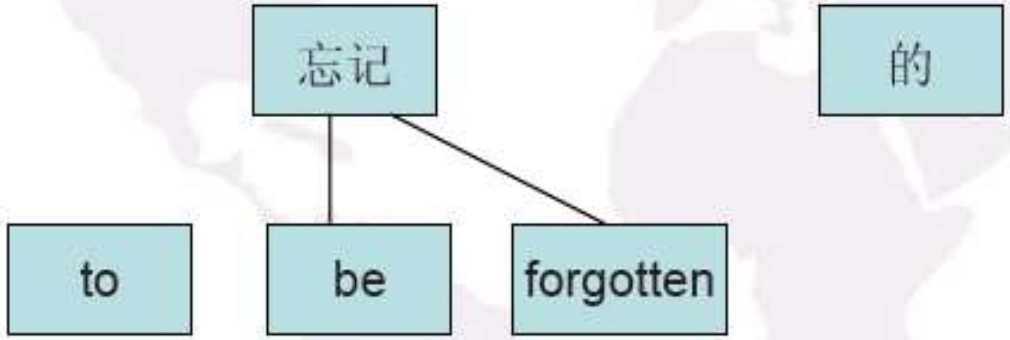
$$p(not \mid 是不能忘记的) = 1$$

对齐错误使得概率计算不准确，影响解码

双语短语的概率计算(4)

引入短语的词汇化翻译概率，利用IBM Model 1训练得到的词语翻译表计算双向的词语翻译概率

$$lex(\tilde{f} \mid \tilde{e}, a) = \prod_{j=1}^{n} \frac{1}{|\{i \mid (j,i) \in a\}|} \sum_{\forall (i,j) \in a} p(f_j \mid e_i)$$



$$lex(\text{忘记 的} \mid to\ be\ forgotten) = \frac{1}{2} \times (p(\text{忘记} \mid be) + p(\text{忘记} \mid forgotten)) \times p(\text{的} \mid NULL)$$

双语短语的概率计算(5)

- 通常都使用双向的短语翻译概率和双向的词汇化翻译概率，一个四个概率作为特征，与其他特征一起，利用最小错误率算法调整特征参数

统计机器翻译的对数线性模型(1)

- Och于ACL2002提出，思想来源于Papineni提出的基于特征的自然语言理解方法，该论文获得ACL2002的最佳论文
- 是一个比信源－信道模型更具一般性的模型，信源－信道模型是其一个特例
- 原始论文的提法是“最大熵”模型，现在通常使用“对数线性（Log-Linear）模型”这个概念
- “对数线性模型”的含义比“最大熵模型”更宽泛，而且现在这个模型通常都不再使用最大熵的方法进行参数训练，因此“对数线性”模型的提法更为准确
- 与NLP中通常使用的最大熵方法的区别：使用连续量（实数）作为特征，而不是使用离散的布尔量（只取0和1值）作为特征

统计机器翻译的对数线性模型(2)

假设e、f是机器翻译的目标语言和源语言句子，$h_1(e,f), \ldots, h_M(e,f)$分别是$e$、$f$上的M个特征，$\lambda_1, \ldots, \lambda_M$是与这些特征分别对应的M个参数，那么翻译概率可以用以下公式模拟：

$$\Pr(e|f) \approx p_{\lambda_1 \ldots \lambda_M}(e|f)$$

$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e,f)]}{\sum_{e'} \exp[\sum_{m=1}^{M} \lambda_m h_m(e',f)]}$$

统计机器翻译的对数线性模型(3)

对于给定的*f*，其最佳译文*e*可以用以下公式表示：

$$\hat{e} = \arg\max_{e}\{\Pr(e \mid f)\}$$

$$\approx \arg\max_{e}\{\sum_{m=1}^{M}\lambda_m h_m(e, f)\}$$

对数线性模型vs.噪声信道模型

- 取以下特征和参数时，对数线性模型等价于噪声信道模型：
  - 仅使用两个特征
  - $h_1(e,f)=\log p(e)$
  - $h_2(e,f)=\log p(f|e)$
  - $\lambda_1=\lambda_2=1$

对数线性模型：Och的实验(1)

- 方案
  - 首先将信源信道模型中的翻译模型换成反向的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降；
  - 调整参数λ1和λ2，系统性能有了较大提高；
  - 再依次引入其他一些特征，系统性能又有了更大的提高。

对数线性模型：Och的实验(2)

- 其他特征
  - 句子长度特征(WP)：对于产生的每一个目标语言单词进行惩罚；
  - 附加的语言模型特征(CLM)：一个基于(词)类的语言模型特征；
  - 词典特征(MX)：计算给定的输入输出句子中有多少词典中存在的共现词对。

对数线性模型：Och的实验(3)

- 实验结果

	objective criteria [%]					subjective criteria [%]	
	SER	WER	PER	mWER	BLEU	SSER	IER
Baseline($\lambda_m = 1$)	86.9	42.8	33.0	37.7	43.9	35.9	39.0
ME	81.7	40.2	28.7	34.6	49.7	32.5	34.8
ME+WP	80.5	38.6	26.9	32.4	54.1	29.9	32.2
ME+WP+CLM	78.1	38.3	26.9	32.1	55.0	29.1	30.9
ME+WP+CLM+MX	77.8	38.4	26.8	31.9	55.2	28.8	30.9

对数线性模型的优点

- 噪声通道模型只有在理想的情况下才能达到最优，对于简化的语言模型和翻译模型，取不同的参数值实际效果更好；
- 对数线性模型大大扩充了统计机器翻译的思路；
- 特征的选择更加灵活，可以引入任何可能有用的特征。
  - 多个翻译模型和语言模型可以同时使用

小结

统计机器翻译的发展可以清理出两条主线的进展

- 框架模型的进展
  - 信源信道模型
  - 对数线性模型
- 翻译模型的进展
  - 基于词的模型
  - 基于短语的模型
  - 基于句法的模型