



中国科学院计算技术研究所

词语对齐的快速增量式训练方法研究

NLP&CC, 北京, 2012年11月4日

罗维

提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

统计机器翻译

▶ 【学术界】 模型朝句法、语义方向发展

- ▶ 基于词
- ▶ 基于短语
- ▶ 基于句法
- ▶ 基于语义



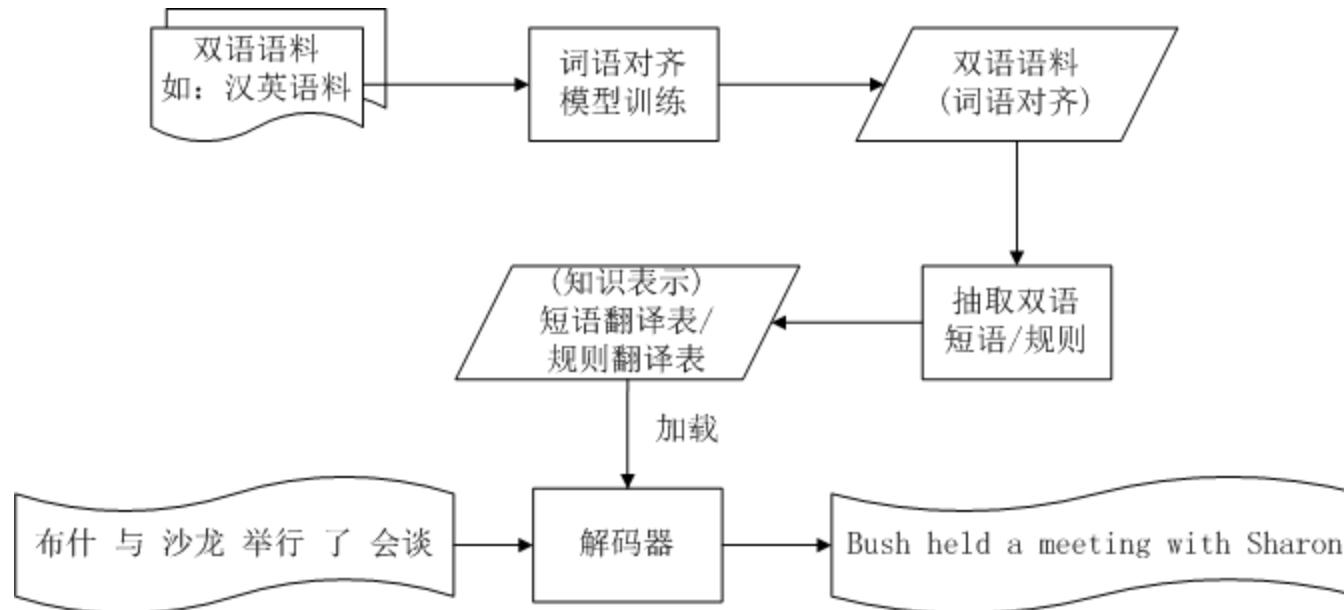
统计机器翻译模型

▶ 【产业界】 正逐渐走向实用

- ▶ 百度翻译、有道翻译、谷歌翻译...

统计翻译模型

► 构建 短语/规则翻译表 的一般流程



► 注：“解码器”模块省略关于翻译特征的判别式训练的内容。

词语对齐

- ▶ 词语对齐是翻译模型构建的瓶颈，这是因为
 - ▶ 主流的翻译模型均依赖大规模双语语料库的词语对齐结果
 - ▶ 词语对齐需要模型训练，在翻译模型构建中耗时最长
 - ▶ 词语对齐质量影响机器翻译质量

词语对齐

- ▶ 词语对齐是翻译模型构建的瓶颈，这是因为
 - ▶ 主流的翻译模型均依赖大规模双语语料库的词语对齐结果
 - ▶ 词语对齐需要模型训练，在翻译模型构建中耗时最长
 - ▶ 词语对齐质量影响机器翻译质量
- ▶ 当新增语料时，传统的模型训练方法是合并语料后重新进行词语对齐模型训练。

词语对齐

- ▶ 词语对齐是翻译模型构建的瓶颈，这是因为
 - ▶ 主流的翻译模型均依赖大规模双语语料库的词语对齐结果
 - ▶ 词语对齐需要模型训练，在翻译模型构建中耗时最长
 - ▶ 词语对齐质量影响机器翻译质量
- ▶ 当新增语料时，传统的模型训练方法是合并语料后重新进行词语对齐模型训练。其两大不足：
 - ▶ 时间和计算资源的开销较大
 - ▶ NIST评测提供的500w平行句对 | 曙光5000的计算节点机 | 单线程 | 4~5天
 - ▶ 当新数据是与原始数据所在领域不同的数据
 - ▶ 传统训练方法得到的模型会与新数据上的真实模型存在较大的偏差。

提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

词语对齐的增量式训练

- ▶ [Wu 2005]
 - ▶ 模型插值
 - ▶ 应用得到的模型在新数据上生成词语对齐结果
- ▶ [Duh 2011]
 - ▶ 对新数据应用贝叶斯估计生成词语对齐矩阵。

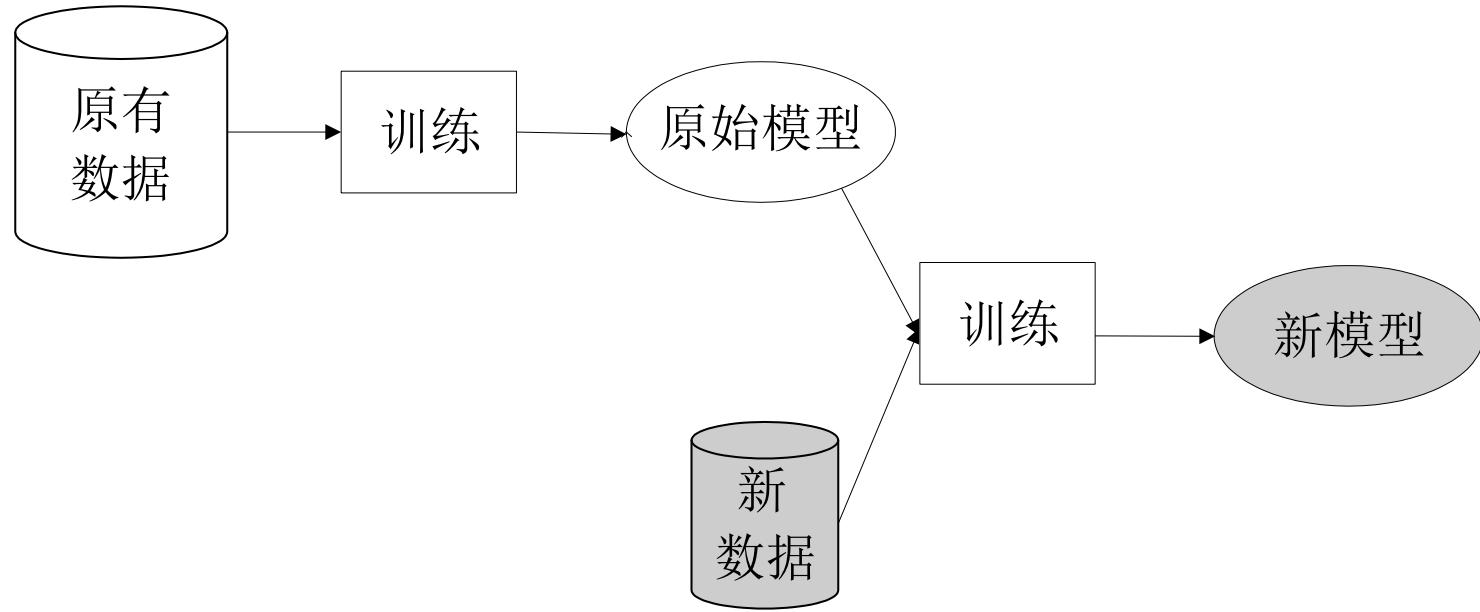
提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ **研究方法**
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

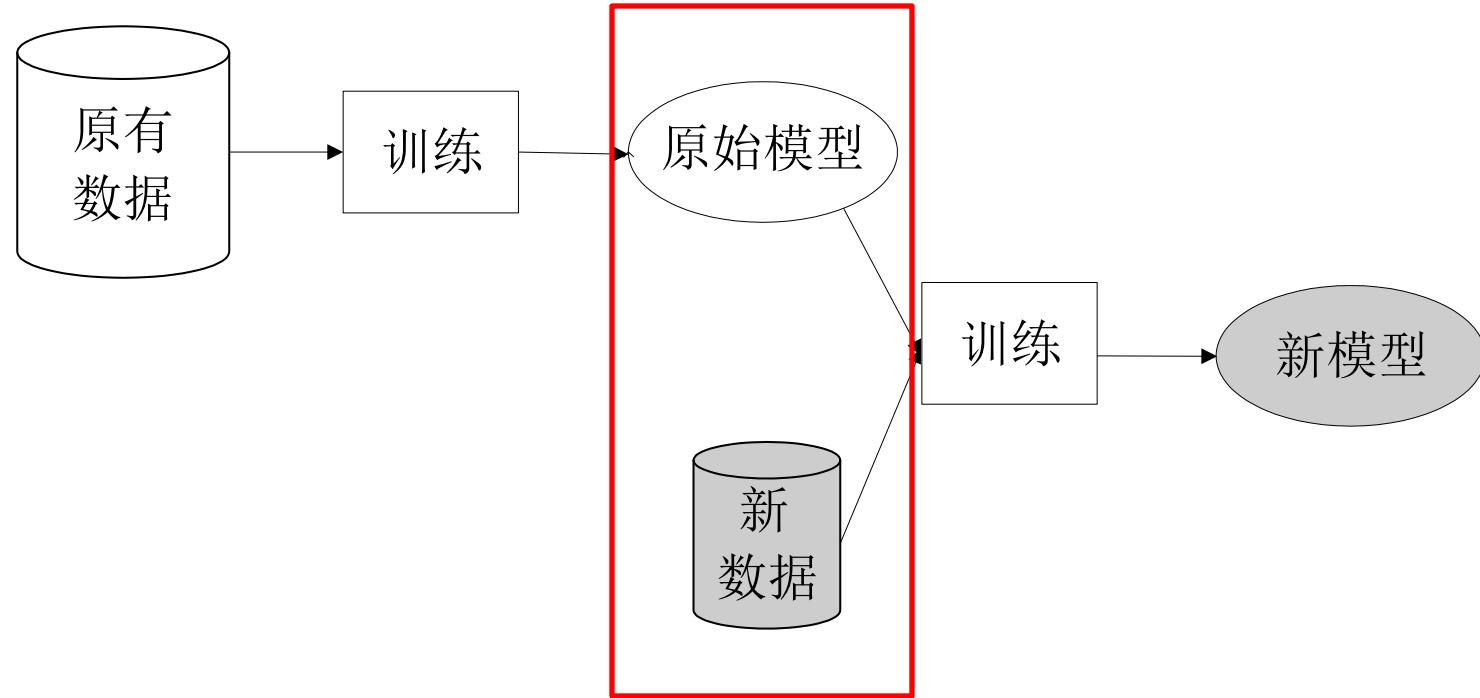
词语对齐模型

- ▶ 采用2个基于无监督学习的模型
 - ▶ IBM模型1 [Brown 93]
 - ▶ 基于HMM模型的词语对齐模型 [Vogel 96]
- ▶ 模型的参数训练算法
 - ▶ EM算法

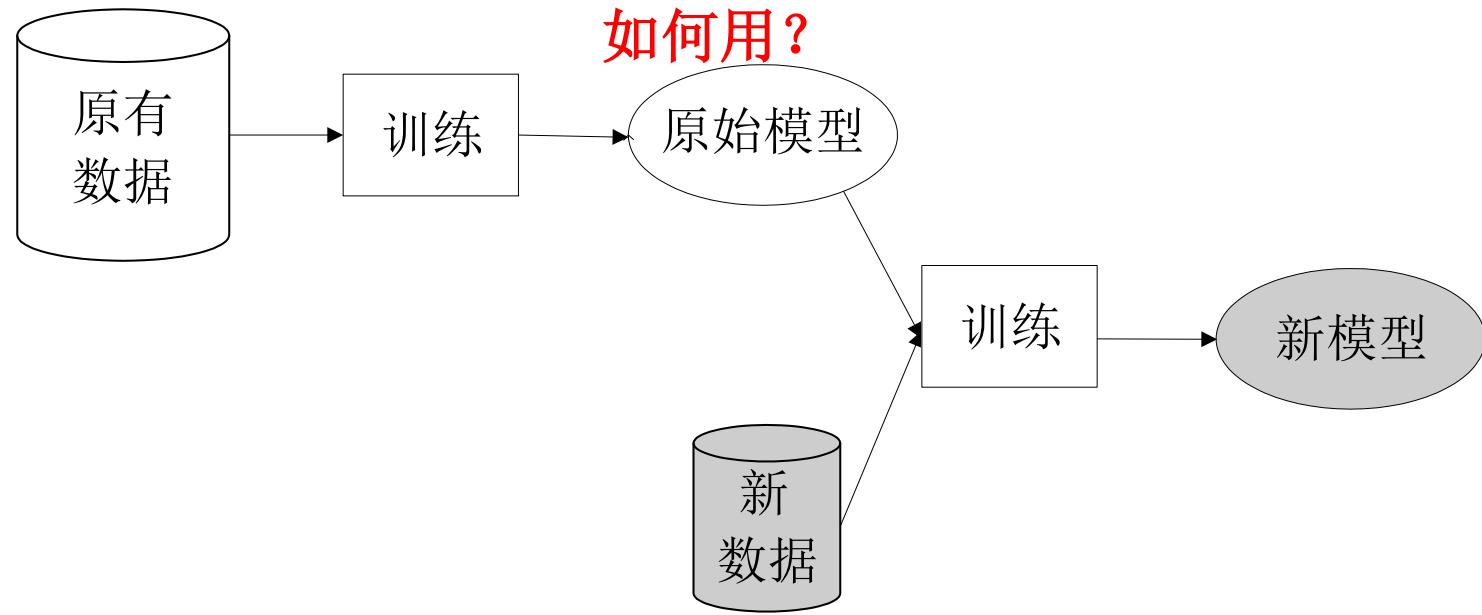
词语对齐模型的增量式训练



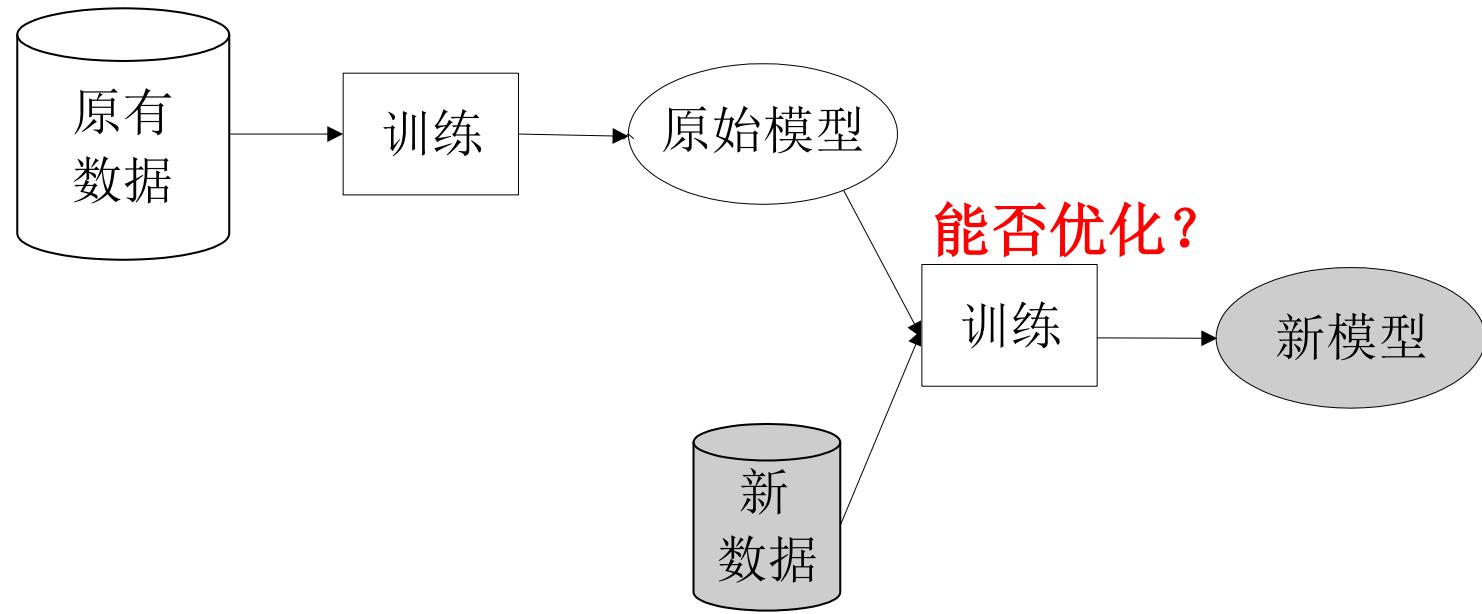
词语对齐模型的增量式训练



词语对齐的增量式训练



词语对齐的增量式训练



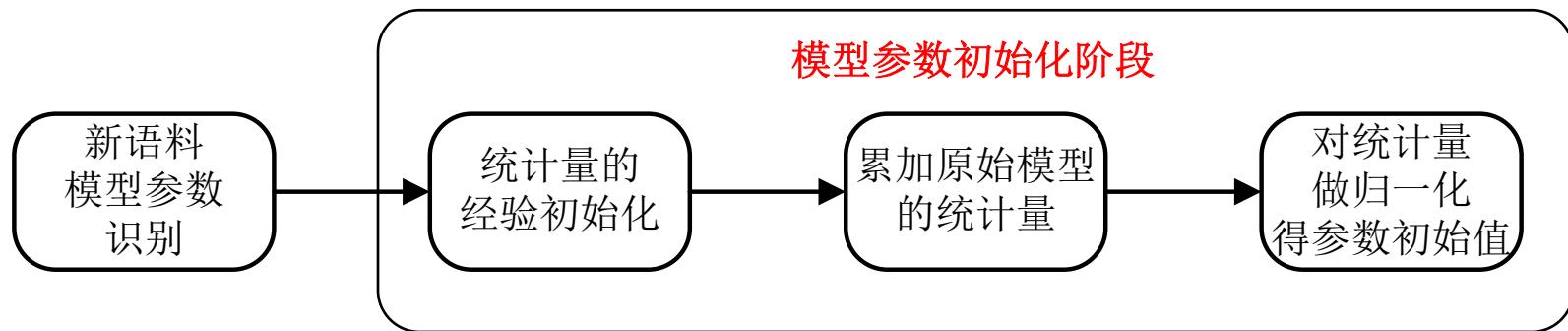
提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

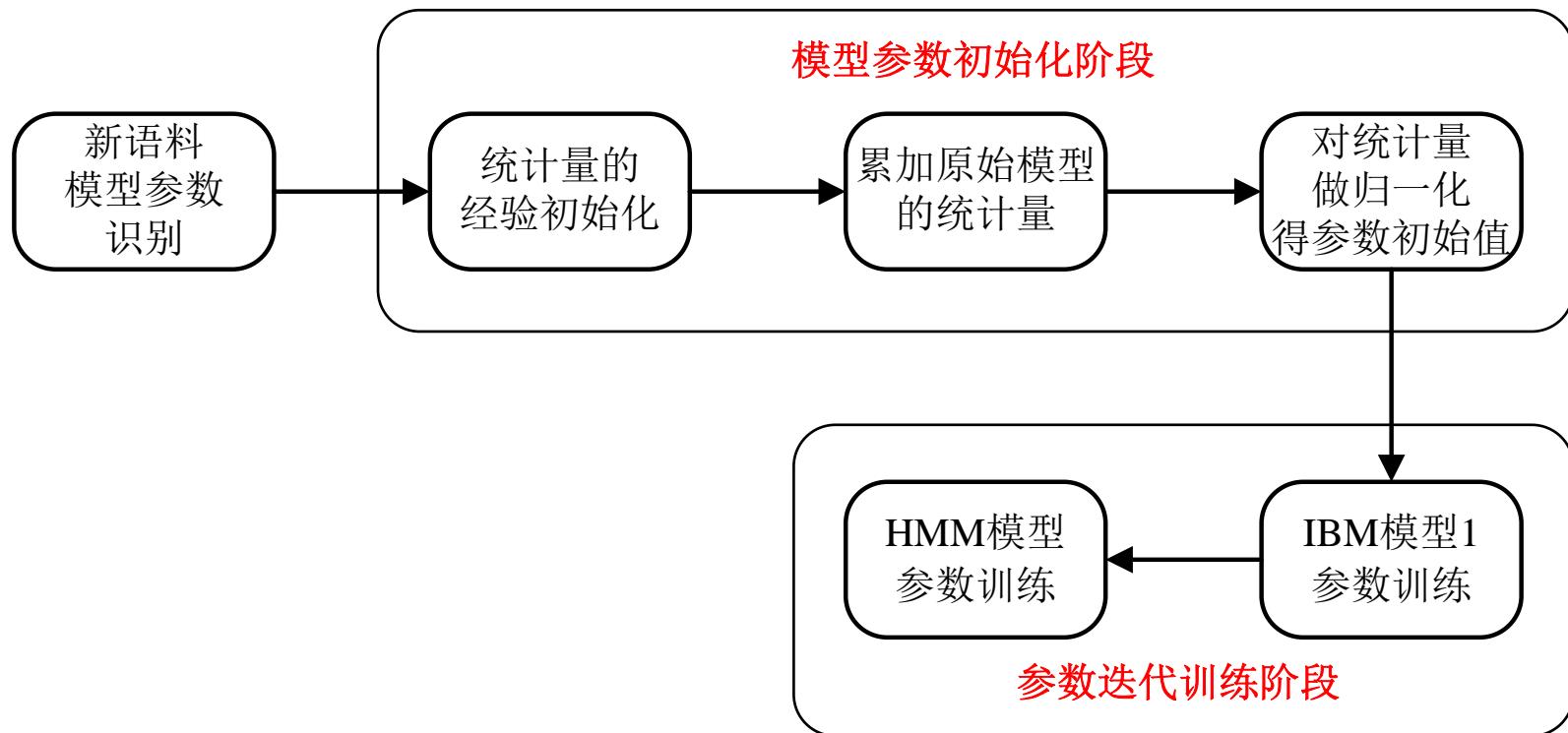
基于初始化的增量式训练

新语料
模型参数
识别

基于初始化的增量式训练



基于初始化的增量式训练



基于初始化的增量式训练

▶ 模型

- ▶ 参数：词汇化翻译概率 $t(f|e)$, 扭曲概率 $d(a_j|a_{j-1}, I)$
- ▶ 参数统计量：频度统计值 $c(f, e)$, $c(a_j, a_{j-1}, I)$

基于初始化的增量式训练

▶ 模型

- ▶ 参数：词汇化翻译概率 $t(f | e)$, 扭曲概率 $d(a_j | a_{j-1}, I)$
- ▶ 参数统计量：频度统计值 $c(f, e)$, $c(a_j, a_{j-1}, I)$
- ▶ 为什么用原始模型的参数统计量来初始化，而不直接用原始模型的参数来初始化？
 - ▶ 不增加计算复杂度。新语料中出现新的模型参数，如 $t(f_{\text{new}} | e)$;
 - ▶ 统一的系统架构。后文的online EM算法，在进行新一轮的迭代时，需要使用现有的统计量。

实验_传统中医

▶ 平行语料

- ▶ 原始数据：来自LDC语料(约125w句对)
- ▶ 新数据：ctzy语料(约224w句对)，选出70w做增量式实验

▶ 语言模型

- ▶ 在ctzy语料英文端，用srilm工具训练英文端5元LM

▶ 开发集

- ▶ ctzy领域数据(861句)，带4个参考译文

▶ 测试集

- ▶ ctzy领域数据(1000句)，带4个参考译文

▶ 运行系统

- ▶ 自己实现的词语对齐系统(Model 1:5轮，HMM: 3轮)
- ▶ 基于层次短语模型的解码器

实验_传统中医

实验方案	开发集	测试集
Baseline	45.78	54.60
Com_old_new	44.18	53.50
Init_lex	44.80	54.03
Init_dist	46.93	56.52
Init_lex_dist	46.43	55.81

实验设置：

Baseline : 直接在新语料上进行模型训练，并生成词语对齐结果。

Com_old_new: 合并原始语料和新语料，然后进行模型训练。

Init_lex : 用原始语料上训出的词汇化翻译概率初始化。

Init_dist : 用原始语料上训出的位置扭曲概率初始化。

Init_lex_dist : 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

实验_传统中医

跨领域不能简单合并语料

实验方案	开发集	测试集
Baseline	45.78	54.60
Com_old_new	44.18	53.50
Init_lex	44.80	54.03
Init_dist	46.93	56.52
Init_lex_dist	46.43	55.81

实验设置：

Baseline : 直接在新语料上进行模型训练，并生成词语对齐结果。

Com_old_new: 合并原始语料和新语料，然后进行模型训练。

Init_lex : 用原始语料上训出的词汇化翻译概率初始化。

Init_dist : 用原始语料上训出的位置扭曲概率初始化。

Init_lex_dist : 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

实验_传统中医

仅初始化词汇化翻译概率方案不好

实验方案	开发集	测试集
Baseline	45.78	54.60
Com_old_new	44.18	53.50
Init_lex	44.80	54.03
Init_dist	46.93	56.52
Init_lex_dist	46.43	55.81

对齐改进，改良规则分数

实验设置：

Baseline : 直接在新语料上进行模型训练，并生成词语对齐结果。

Com_old_new: 合并原始语料和新语料，然后进行模型训练。

Init_lex : 用原始语料上训出的词汇化翻译概率初始化。

Init_dist : 用原始语料上训出的位置扭曲概率初始化。

Init_lex_dist : 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

实验_医药

▶ 平行语料

- ▶ 原始语料：来自LDC语料(约125w句对)
- ▶ 新语料：yiyao语料(约525w句对)，选出70w做增量式实验

▶ 语言模型

- ▶ 在yiyao语料英文端，用srilm工具训练英文端5元LM

▶ 开发集

- ▶ yiyao领域数据(1024句)，带4个参考译文

▶ 测试集

- ▶ yiyao领域数据(1000句)，带4个参考译文

▶ 运行系统

- ▶ 自己实现的词语对齐系统(Model 1:5轮，HMM: 3轮)
- ▶ 基于层次短语模型的解码器

实验_医药

实验方案	开发集	测试集
Baseline	18.73	38.72
Com_old_new	19.08	38.64
Init_lex	18.89	38.52
Init_dist	19.14	40.74
Init_lex_dist	19.27	40.80

实验设置：

Baseline : 直接在新语料上进行模型训练，并生成词语对齐结果。

Com_old_new: 合并原始语料和新语料，然后进行模型训练。

Init_lex : 用原始语料上训出的词汇化翻译概率初始化。

Init_dist : 用原始语料上训出的位置扭曲概率初始化。

Init_lex_dist : 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

跨领域不能简单合并语料

实验_医药

实验方案	开发集	测试集
Baseline	18.73	38.72
Com_old_new	19.08	38.64
Init_lex	18.89	38.52
Init_dist	19.14	40.74
Init_lex_dist	19.27	40.80

仅初始化词汇化翻译概率方案不好

对齐改进，改良规则分数

实验设置：

Baseline : 直接在新语料上进行模型训练，并生成词语对齐结果。

Com_old_new: 合并原始语料和新语料，然后进行模型训练。

Init_lex : 用原始语料上训出的词汇化翻译概率初始化。

Init_dist : 用原始语料上训出的位置扭曲概率初始化。

Init_lex_dist : 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

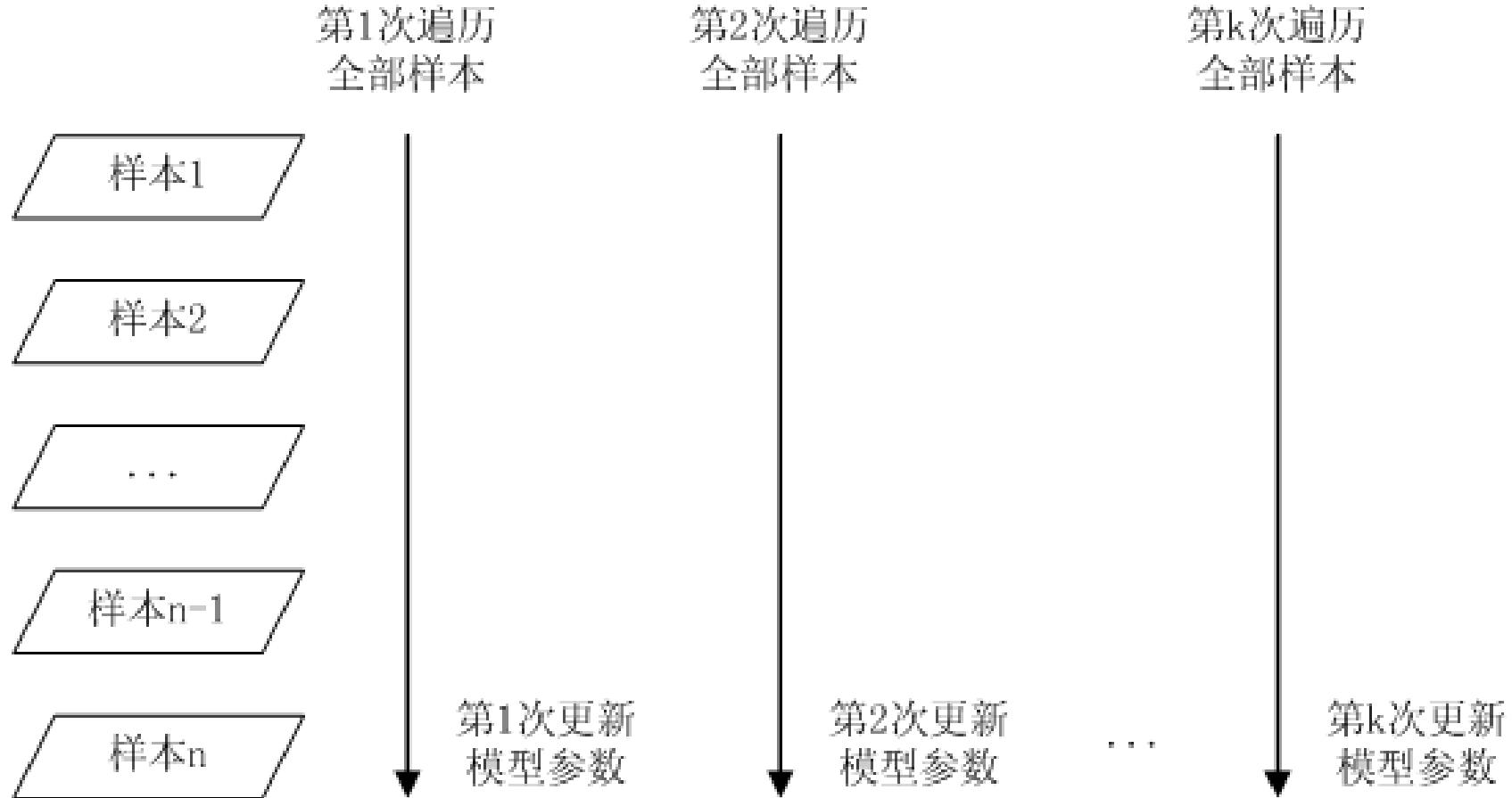
提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

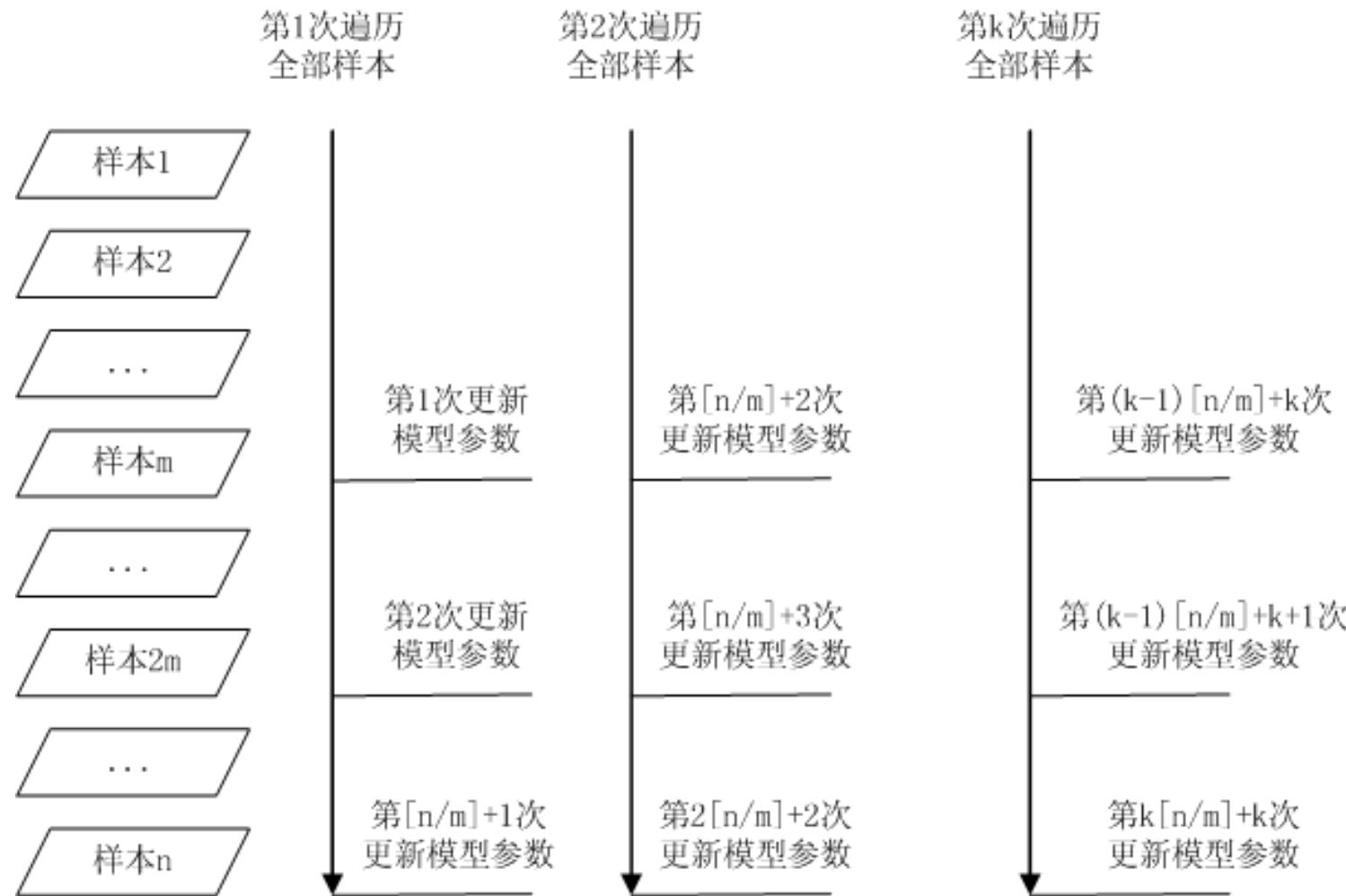
Online EM算法的应用

- ▶ EM算法家族
 - ▶ batch EM (属于batch learning)
 - ▶ online EM (属于online learning)

Batch learning



Online learning



Online learning

- ▶ 与Batch learning方法，online算法的优点：
 - ▶ 容易跳出差的局部最优解；
 - ▶ 在大数据量的参数优化的效果较好；
 - ▶ 如果训练数据中存在冗余，能更快地到达（局部）最优解。
- ▶ Online算法缺点：
 - ▶ 参数更新频繁，增加了运算量
 - ▶ 并行程度不会有batch算法高

Online learning

- ▶ 与Batch learning方法，online算法的优点：
 - ▶ 容易跳出差的局部最优解；
 - ▶ 在大数据量的参数优化的效果较好；
 - ▶ 如果训练数据中存在冗余，能更快地到达（局部）最优解。
- ▶ Online算法缺点：
 - ▶ 参数更新频繁，增加了运算量
 - ▶ 并行程度不会有batch算法高

这优缺点的对比分析自然也适用于batch EM和online EM的对比分析。

Batch EM

[Liang 2009]

```
 $\mu \leftarrow$  initialization
for each iteration  $t = 1, \dots, T$ 
   $\mu' \leftarrow 0$ 
  for each example  $i = 1, \dots, n$ :
     $s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$  [inference]
     $\mu' \leftarrow \mu' + s_i'$  [accumulate new]
   $\mu \leftarrow \mu'$  [replace old with new]
```

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Batch EM

μ 表示由所有充分统计量组成的一个向量

$\mu \leftarrow$ initialization

for each iteration $t = 1, \dots, T$

$\mu' \leftarrow 0$

for each example $i = 1, \dots, n$:

$s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$ [inference]

$\mu' \leftarrow \mu' + s_i'$ [accumulate new]

$\mu \leftarrow \mu'$ [replace old with new]

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Batch EM

```
 $\mu \leftarrow \text{initialization}$ 
for each iteration  $t = 1, \dots, T$ 
   $\mu' \leftarrow 0$ 
  for each example  $i = 1, \dots, n$ :
     $s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$  [inference]
     $\mu' \leftarrow \mu' + s_i'$  [accumulate new]
   $\mu \leftarrow \mu'$  [replace old with new]
```

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

E步

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Batch EM

```
 $\mu \leftarrow \text{initialization}$ 
for each iteration  $t = 1, \dots, T$ 
   $\mu' \leftarrow 0$ 
  for each example  $i = 1, \dots, n$ :
     $s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$  [inference]
     $\mu' \leftarrow \mu' + s_i'$  [accumulate new]
   $\mu \leftarrow \mu'$  [replace old with new]
```

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

M步

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Online EM – stepwise EM

[Liang 2009]

$\mu \leftarrow \text{initialization}; k = 0$

for each iteration $t = 1, \dots, T$

for each example $i = 1, \dots, n$:

$s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$ [inference]

$\mu \leftarrow (1 - \eta_k)\mu + \eta_k s_i'; k \leftarrow k + 1$ [towards new]

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Online EM – stepwise EM

$\mu \leftarrow \text{initialization}; k = 0$

for each iteration $t = 1, \dots, T$

for each example $i = 1, \dots, n$:

$s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$ [inference]

$\mu \leftarrow (1 - \eta_k) \mu + \eta_k s_i'; k \leftarrow k + 1$ [towards new]

E步

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Online EM – stepwise EM

$\mu \leftarrow \text{initialization}; k = 0$

for each iteration $t = 1, \dots, T$

for each example $i = 1, \dots, n$:

$$s'_i \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$$

[inference]

$$\mu \leftarrow (1 - \eta_k) \mu + \eta_k s'_i; k \leftarrow k + 1$$

[towards new]

M步

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s'_i : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Online EM – stepwise EM

$\mu \leftarrow \text{initialization}; k = 0$

for each iteration $t = 1, \dots, T$

for each example $i = 1, \dots, n$:

$$s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z) \quad [\text{inference}]$$

$$\mu \leftarrow (1 - \eta_k) \mu + \eta_k s_i'; k \leftarrow k + 1 \quad [\text{towards new}]$$

与batch EM的这步一样

z : 与 x 对应的隐变量

μ : 模型参数对应的充分统计量

s_i' : 从样本 i 得到的充分统计量

$\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数

$\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

Online EM – stepwise EM

```
 $\mu \leftarrow \text{initialization}; k = 0$ 
for each iteration  $t = 1, \dots, T$ 
  for each example  $i = 1, \dots, n$ :
     $s_i' \leftarrow \sum_z p(z | x^{(i)}; \Theta(\mu)) \Phi(x^{(i)}, z)$  [inference]
     $\mu \leftarrow (1 - \eta_k) \mu + \eta_k s_i'; k \leftarrow k + 1$  [towards new]
```



z : 与 x 对应的隐变量
 μ : 模型参数对应的充分统计量
 s_i' : 从样本 i 得到的充分统计量
 $\Theta(\mu)$: 给定统计量 μ , 依据最大似然估计计算模型参数
 $\Phi(x, z)$: 标注数据 (x, z) 下映射出的充分统计量

每轮迭代完后，不用清空充分统计量

Online EM – stepwise EM

- ▶ η_k 的取值

- ▶ EM 收敛要求 $\sum_{k=0}^{\infty} \eta_k = \infty$ 与 $\sum_{k=0}^{\infty} \eta_k^2 < \infty$
- ▶ $\eta_k = (k + 2)^{-\alpha}$ ($0.5 < \alpha \leq 1$)

- ▶ mini-batch 的取值

- ▶ 每处理 mini-batch 个样本，更新模型参数

实验_Online EM算法

- ▶ ctzy 实验
 - ▶ 与之前ctzy实验配置一样
- ▶ yiya 实验
 - ▶ 与之前yiya实验配置一样
- ▶ 运行系统
 - ▶ 自己实现的词语对齐系统
 - ▶ model 1: online EM训练3轮
 - ▶ hmm : online EM训练3轮
 - ▶ 基于层次短语模型的解码器

实验_Online EM算法—传统中医

实验方案	batch EM		online EM	
	开发集	测试集	开发集	测试集
Baseline	45.78	54.60	X	X
Init_lex_dist	46.43	55.81	46.68	56.37

实验设置：

Baseline：直接在新语料上进行模型训练，并生成词语对齐结果。

Init_lex_dist：用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

实验_Online EM算法—医药

实验方案	batch EM		online EM	
	开发集	测试集	开发集	测试集
Baseline	18.73	38.72	X	X
Init_lex_dist	19.27	40.80	18.96	40.82

实验设置：

Baseline：直接在新语料上进行模型训练，并生成词语对齐结果。

Init_lex_dist: 用原始语料上训出的词汇化翻译概率和扭曲概率初始化。

提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ **总结**
- ▶ 主要参考文献

总结

- ▶ 传统训练方法在构建翻译模型时存在不足，本文围绕翻译模型构建流程的瓶颈——词语对齐，研究既高效又保证词语对齐质量和机器翻译质量的增量式训练方法。

总结

- ▶ 传统训练方法在构建翻译模型时存在不足，本文围绕翻译模型构建流程的瓶颈——词语对齐，研究既高效又保证词语对齐质量和机器翻译质量的增量式训练方法。
- ▶ 提出了基于初始化，同时应用迭代训练收敛速度更快的online EM算法以替换通常所用的batch EM算法的增量式训练方法。

总结

- ▶ 传统训练方法在构建翻译模型时存在不足，本文围绕翻译模型构建流程的瓶颈——词语对齐，研究既高效又保证词语对齐质量和机器翻译质量的增量式训练方法。
- ▶ 提出了基于初始化，同时应用迭代训练收敛速度更快的online EM算法以替换通常所用的batch EM算法的增量式训练方法。
- ▶ 所提出的增量式训练方法在实验中取得成效。

提纲

- ▶ 研究背景
- ▶ 研究现状
- ▶ 研究方法
- ▶ 论文的主要工作
 - ▶ 基于初始化的增量式训练
 - ▶ Online EM 算法的应用
- ▶ 总结
- ▶ 主要参考文献

主要参考文献

- ▶ [Brown 1993] Brown, P.F. and Pietra, V.J.D. and Pietra, S.A.D. and Mercer, R.L. The mathematics of statistical machine translation: Parameter estimation. CL 1993.
- ▶ [Duh 2011] Duh, K. and Sudoh, K. and Iwata, T. and Tsukada, H. Alignment Inference and Bayesian Adaptation for Machine Translation. MTSummit 2011.
- ▶ [Liang 2009] Liang, P. and Klein, D. Online EM for unsupervised models. NAACL 2009.
- ▶ [Vogel 1996] Vogel, S. and Ney, H. and Tillmann, C. HMM-based word alignment in statistical translation. ACL 1996.
- ▶ [Wu 2005] Hua, W. and Haifeng, W. and Zhanyi, L. Alignment model adaptation for domain-specific word alignment. ACL 2005.



中国科学院计算技术研究所

谢谢！



IBM词语对齐模型1

[Brown 1993]

► 优化目标：

$$\Pr(F | E) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^l \dots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j | e_{a_j})$$

- E: 目标端句子
- l: 目标端句子长度
- e_i : 目标端句子中位于i位置的词语
- F: 源端句子
- m: 源端句子长度
- f_j : 源端句子中位于j位置的词语
- a_j : 源端j位置对应到目标端的位置

IBM词语对齐模型1

[Brown 1993]

▶ 优化目标：

$$\Pr(F | E) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^l \dots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j | e_{a_j})$$

▶ EM算法：

▶ E步

$$c(f | e; F, E) = \sum_A \Pr(A | E, F) \sum_{j=1}^m \delta(f, f_j) \delta(e, e_{a_j})$$

▶ M步

$$t(f | e) = \lambda_e^{-1} c(f | e; F, E)$$

aj : 视为隐变量
t(fj | ei): 模型参数

基于HMM模型的词语对齐

[Vogel 1996]

▶ 优化目标：

$$\Pr(F | E) = \sum_{a_1=0}^l \dots \sum_{a_m=0}^l \prod_{j=1}^m p(a_j | a_{j-1}, l) t(f_j | e_{a_j})$$

▶ EM 算法：

▶ 前向后向算法

a_j : 视为隐变量

$t(f_j | e_i)$: 模型参数

$p(a_j | a_{j-1}, l)$: 模型参数

Online EM – stepwise EM

► Sparse update 问题

- 处理一个或几个样本时，能够影响到的统计量只局限于某几个统计量上。
- $\mu \leftarrow (1 - \eta_k)\mu + \eta_k s_i'$



Online EM – stepwise EM

► Sparse update 问题

- 处理一个或几个样本时，能够影响到的统计量只局限于某几个统计量上。
- $\mu \leftarrow (1 - \eta_k)\mu + \eta_k s_i'$

► 通过引入

$$S = \frac{\mu}{\prod_{j < k} (1 - \eta_j)}$$

► 统计量的插值计算公式等价于

$$S \leftarrow S + \frac{\eta_k}{\prod_{j \leq k} (1 - \eta_j)} s_i'$$

