北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

词语对齐的快速增量式训练方法研究

罗维

中国科学院计算技术研究所，北京 100190; † E-mail: vividfree@gmail.com

摘要 围绕翻译模型构建流程的瓶颈——词语对齐，着手翻译模型的增量式训练。在基于无监督学习的词语对齐模型的基础上，提出一种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 算法的方法，实现增量式训练。实验表明，所提出的方法既高效又能保证词语对齐质量和机器翻译质量。

关键词 统计机器翻译；词语对齐；增量式训练；期望最大化；在线算法

中图分类号 TP391

Research on Fast Incremental Training Algorithm for Word Alignment

LUO Wei

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190; † E-mail: vividfree@gmail.com

Abstract This study puts emphasis on the incremental training algorithm for word alignment, which is the bottleneck during the construction of translation model. Based on two unsupervised word alignment models, the authors propose an incremental training algorithm which is based on initialization and online EM algorithm. Experiments show that the proposed methods are efficient and would not hurt the quality of word alignment and translation.

Key words statistical machine translation; word alignment; incremental training; expectation maximization; online algorithm

近年来，统计机器翻译取得了很大的成功。统计机器翻译模型依赖于大规模双语语料库的训练。当新增语料时，传统的翻译模型训练方法需要首先合并新语料和原始语料，并重新进行模型训练。这种训练方法存在两方面不足。其一，当训练数据规模增大到一定规模时，这种训练方法需要消耗大量的时间和计算资源。其二，如果新数据是与原始数据所在领域不同的数据，那么该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。研究既高效又保证机器翻译质量的翻译模型增量式训练具有重要的研究意义和实用价值。

由于翻译模型的构建一般起始于词语对齐，耗时最长的阶段也在于词语对齐，而且词语对齐的质量会最终影响机器翻译质量，因而词语对齐成为翻译模型训练的一个瓶颈。然而，目前只有很少的文献对词语对齐的增量式训练做了探讨[1−2]。文献[1]对原始数据上训练出的原始模型和新数据上训练出的新模型进行插值得到最终的模型，再在新数据上生成最终的词语对齐结果。文献[2]则应用已训练得到的原始模型和新模型，对新数据应用贝叶斯估计生成词语对齐矩阵。本文与文献[1]和[2]所采用的研究方法不同，提出一种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 算法的方法，实现增量式训练。实验表明，所提出的增量式训练方法既高效又保证词语对齐质量和机器翻译质量。

863 计划(2011AA01A207)资助
收稿日期: 2012−06−05; 修回日期: 2012−08−30; 网络出版时间: 2012-10-26 17:55
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.019.html

1 词语对齐模型

词语对齐是翻译模型构建的首要一步，在经过词语对齐模型训练后，生成双语语料句对的词语对齐结果。机器翻译解码器的重要知识源——短语翻译表或者规则翻译表，一般都是从双语语料的词语对齐结果中抽取得到。词语对齐质量，影响着短语翻译表和规则翻译表，并最终影响着机器翻译质量。

用 $F = f_1^J = f_1 f_2 ... f_J$ 表示源语言句子，J 为其长度，$E = e_1^I = e_1 e_2 ... e_I$ 表示目标语言句子，I 为其长度。e_i 和 f_j 互为翻译(或者部分翻译)，记为 $a_j = i$，或者用元组记为 $l = (i, j)$。对齐 A 被定义为词语位置的笛卡儿集的子集：$A \subseteq \{(i, j): \quad i = 1,...,I; \ j = 1,...,J\}$。给定平行句对语料库 $\{\boldsymbol{F}, \boldsymbol{E}\}$，其中 $\boldsymbol{F}$ 表示源语言句子的集合，而 $\boldsymbol{E}$ 表示由与每个源语言句子对应的目标语言句子所组成的集合。我们希望从中找到每个句对 $(F, E) \in \{\boldsymbol{F}, \boldsymbol{E}\}$ 之间最有可能的词语对齐关系 A。

本文在两个基于无监督学习的词语对齐模型——IBM 模型 1[3]以及基于 HMM 的词语对齐模型[4]的基础上，研究增量式训练。IBM 模型 1 只有词语化翻译模型，而基于 HMM 的词语对齐模型则比 IBM 模型 1 多了扭曲模型。经过 IBM 模型 1 训练得到的词汇化翻译概率将作为基于 HMM 的词语对齐模型中词汇化翻译概率的初始值。尽管还有一些更复杂的模型，如 IBM 模型 3 到 6[3,5]，但是学者一般认为经过基于 HMM 的词语对齐模型训练得到的词语对齐质量与那些复杂模型的对齐质量相差不多[6]。当然，这并不影响将下文所提出的词语对齐增量式训练方法延伸到 IBM 模型 3 到 6 中。

对于基于 HMM 的词语对齐模型，其针对对齐位置的跳转做了 1 阶马尔科夫假设，并进行了如下建模:

$$\Pr(F, A \mid E) = \Pr(J \mid E) \prod_{j=1}^{J} (d(a_j \mid a_{j-1}, I) \times t(f_j \mid e_{a_j})), \quad (1)$$

其中 $t(f_j \mid e_{a_j})$ 是词汇化翻译模型，$d(a_j \mid a_{j-1}, I)$ 是扭曲模型。通常而言，IBM 模型 1 和基于 HMM 的词语对齐模型的参数训练，均是在似然函数最大的优化目标下，应用 EM 算法[3, 8]来得到。

2 词语对齐的增量式训练

图 1 描述了本文增量式训练方法的大致思路，并不是先合并原始语料和新语料后再重新训练模型参数，而是通过某种方法，在原始模型的基础上，实现模型的增量式训练。如果能够充分利用已经训练好的原有模型，仅对新的数据重新训练，则可以大大提高训练速度，节省训练资源，同时还有可能得到与真实模型偏差(bias)更小的模型。倘若该思路应用于词语对齐的增量式训练，那么图 1 所涉及的模型均是指词语对齐模型。对于 IBM 模型 1 和基于 HMM 的词语对齐模型，EM 算法是这两个模型的经典参数训练算法。

围绕图 1 所示的增量式训练方法思路，下文将针对模型的参数训练算法，先后应用两个方法，研究既高效又保证词语对齐和机器翻译质量的增量式训练方法。

2.1 基于初始化的增量式训练

图 2 描述了所提出的基于初始化的增量式训练算法流程。首先，通过经验取值，给新模型中所有参数对应的充分统计量①进行初始化；其次，把原始模型的充分统计量加到新模型的充分统计量上；

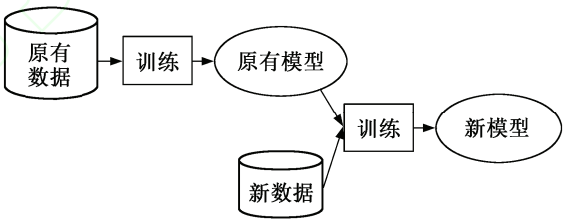

图 1 新数据到来时增量式训练方法示意图

Fig. 1 Diagram of incremental training algorithm for new coming data

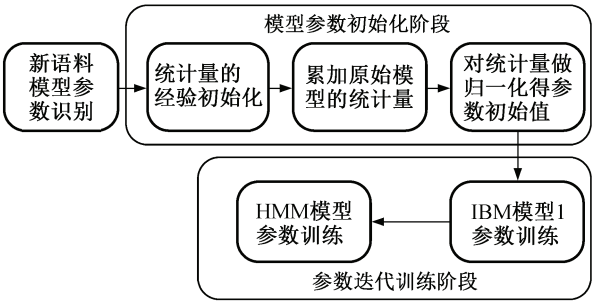

图 2 基于初始化的词语对齐增量式训练算法流程图

Fig. 2 Procedure of initialization based incremental training algorithm for word alignment

① 对于多项式分布来说，模型的充分统计量是指一系列的频度统计 count。例如，IBM 模型 1 中的词汇化翻译模型是由很多词语对的词汇化翻译概率 t(f | e)组成，而每个 t(f | e)所对应的充分统计量就是源端词语 f 和目标端词语 e 共现的频度 c(f, e)。

继而在整体上进行一次归一化操作，得到新模型的参数初始值；最后使用之前介绍的 IBM 模型 1 和 HMM 模型的 EM 算法，经过若干轮迭代训练，得到最终的新模型参数。

这种方法不舍弃原始模型，而是借助原始模型的信息来初始化新模型，随后在新数据上执行词语对齐模型的参数训练，有利于得到更适合新语料的参数估计，并且不会给模型参数的迭代训练阶段增加计算复杂度。同时它还适合不能提供原始数据场景下词语对齐的增量式训练，而合并新旧语料后再进行的传统训练方法对该场景将无能为力。在实际中，不能提供原始数据的场景是时常碰到的，例如：1) 国家安全部门、军方等单位(甲方)不能与技术提供方(乙方)互相交换数据；2) 机器翻译系统的用户希望自己能独立进行翻译模型改进的训练。

注意到，所提出的增量式训练算法在模型参数初始化阶段使用的是原始模型的参数统计量，为什么不直接用原始模型的参数来初始化新模型的参数呢？有两方面的原因。首先，所提出的方法不影响计算效率。即使是直接用模型参数来初始化，对于新语料含有的新模型参数，需要进行初始化赋值(否则 EM 算法将会造成该模型参数一直为 0)，这时还是需要一次归一化操作，所以所提出的方法不影响计算效率。其次，2.2 节将要介绍的 online EM 算法在进行模型的新一轮迭代训练时，需要维持现有的统计量，而不是清空，为了使整个词语对齐系统在使用 batch EM 算法或者 online EM 算法时保持一个统一的计算框架，我们采用了初始化模型参数的统计量这一思路。

当然，在模型参数初始化阶段中，先对统计量做初始化，设初始化值为 a，然后再加上原始模型的统计量。对 a 的不同取值，就相当于在原始模型统计量和新模型统计量上进行的插值。a 取值越小，就越侧重于原始模型的统计量；a 取值越大，就越不侧重于原始模型的统计量。

由于 IBM 模型 1 和基于 HMM 的词语对齐模型总共含有两类模型参数：词汇化翻译概率和扭曲概率。所以在累加原始模型的充分统计量一步中，就可以有如下 4 种考虑：1) 不用原始模型中的任何统计量来初始化；2) 只用原始模型的词汇化翻译概率对应的统计量来初始化；3) 只用原始模型的扭曲概率对应的统计量来初始化；4) 原始模型的词汇化翻译概率和扭曲概率对应的统计量都用来初始化。

2.2 Online EM 算法的应用

传统的 IBM 模型 1 和基于 HMM 的词语对齐模型在进行参数训练时使用的是 batch EM 算法。该算法在每次遍历全部数据后才更新一次模型参数。参数更新慢，使得模型的迭代收敛速度不快。对此，文献[7]引入 online EM 算法于词语对齐模型训练中，不过并没有在大规模数据上研究应用 online EM 算法训练生成的词语对齐模型和词语对齐结果对机器翻译质量的影响。

Online EM 算法每处理若干个样本就更新一次模型参数，参数更新速度快，迭代收敛速度也快。由于文献[7]通过实验说明 stepwise EM 算法是其中性能较好的 online EM 算法的变种，因而本文将应用 stepwise EM 算法，使得在增量式训练中，模型参数训练速度更快。

算法 1 Stepwise EM 算法框架

Input: mini-batches of sample collection$\{ M : M \subset \{X\} \}$
Input: mini-batch size m
Input: stepsize reduction power α
Output: MLE $\hat{\theta}$
  $\mu \leftarrow$ initialization
  $k = 0$
  for iteration $t = 1,\ldots,T$:
    foreach mini-batch $m \subset M$:
      $\mu' \leftarrow 0$
      foreach $x \in m$:
        $s_i' = \sum_z p(z \mid x;\theta(\mu))\phi(x,z)$ [inference]
        $\mu' \leftarrow \mu' + s_i'$
      $\eta_k = (k+2)^{-\alpha}$
      $\mu \leftarrow (1-\eta_k) * \mu + \eta_k * \mu'$ [towards new]
      $k \leftarrow k + 1$

文献[7]给出了 stepwise EM 算法的框架，不过描述的只是每处理一个样本就更新一次参数权重的情形。为了达到每批量处理 m 个样本才更新一次参数权重的目的，我们对框架进行修改。算法 1 用伪代码介绍了 stepwise EM 算法，其中 x 表示观测变量，z 表示与观测变量相对应的隐藏变量，θ 表示模型的参数，$p(x,z;\theta)$ 表示给定参数下 θ，x 和 z 的联合概率分布，$\phi(x,z)$ 表示从完整的样本(x, z)到模型参数的充分统计量的映射，s_i 表示从样本 x 中计算出的充分统计量的期望。$\theta(\mu)$ 表示在 EM 算法的 M 步中，对从样本中收集到的充分统计量的向量 μ，应用最大似然估计方法求解得到新的模型参数。在词语对齐模型中，由于模型参数都是多项式分布，所以 $\theta(\mu)$ 等价于用归一化技术来求解出新的模型参数。

需要注意的是，步长 η_k 依赖于批量处理的规模 m。当观察和处理了越来越多的新数据时，当前的模型参数已经越来越接近真实的参数分布，所以赋给从新观测到的数据上计算出的统计量的权重系数将越来越小。另外，依据文献[7]，设置步长 η_k 与步长衰减指数 α 之间的关系为 $\eta_k=(k+2)^{-\alpha}$，其中 $0.5<\alpha\leqslant 1$，这能保证 online EM 算法收敛到局部最优解。α 越小，步长 η_k 就越大，使得原来的充分统计量的衰减速度更快，也就会使得调整参数的变化幅度更大，但是使得算法的不稳定性会增加。当 α 固定时，随着步骤数 k 的增加，η_k 的值将逐渐变小，使得参数统计量的更新幅度将越来越小。

3 实验

上文针对模型的参数训练算法，先后应用两个方法，研究了词语对齐的增量式训练算法。本节针对原始数据来自于通用领域，而新数据来自于特定领域的情形，实施有关的实验。在汉英方向的机器翻译结果上进行有关的对比实验，并对结果进行分析。

IBM 模型 1 和基于 HMM 的词语对齐模型均是一对多的对齐模型[5]，为达到对齐为多对多的目的，需要首先训练出两个单向(即汉英方向和英汉方向)的原始词语对齐模型，继而在平行句对上生成两种词语对齐结果，再应用启发式策略得到最终的对齐结果。由于词语对齐模型包括词汇化翻译模型和扭曲模型，所以实际中基于初始化的增量式训练最多会涉及如下 4 个模型：汉英方向的通用领域词汇化翻译模型、汉英方向的通用领域扭曲模型、英汉方向的通用领域词汇化翻译模型和英汉方向的通用领域扭曲模型。

3.1 实验配置

首先，从 LDC 数据中提取约 125 万平行句对[①]作为通用领域的数据，从由北京东方灵盾科技有限公司提供的传统中医领域语料中提取约 224 万平行句对作为传统中医领域实验的平行句对，并从中选择 70 万平行句对作为增量式训练的特定领域的数据；并同样从该公司提供的医药领域语料中提取约 525 万平行句对作为医药领域实验的平行句对，也从中选择 70 万平行句对作为增量式训练的特定领域的数据。

目前尚没有官方公布的传统中医领域或者医药领域的开发集和测试集。为了帮助解码器能够自动调整特征权重以及保证对比实验的进行，从北京东方灵盾科技有限公司提供的传统中医领域语料中提取源语言端(即汉语端)句子出现 4 次而目标端(即英语端)句子不同的平行句对，作为传统中医领域实验的开发集(共含有 861 句)和测试集(共含有 1000 句)[②]。与传统中医领域实验抽取开发集和测试集的思路相同，也抽取出医药领域实验的开发集(共含有 1024 句)和测试集(共含有 1000 句)。以大小写不敏感的 NIST BLEU[10]作为翻译质量的评价指标。

用自己研发的词语对齐系统对平行语料进行词语对齐，使用 grow-diag-final 启发式策略[12]来融合两个单向的词语对齐结果。其中，IBM 模型 1 的默认迭代轮数为 5 轮，基于 HMM 的词语对齐模型的默认迭代轮数为 3 轮。如果没有具体说明这两个词语对齐模型的迭代轮数，则按照默认的迭代轮数进行参数迭代训练。在参数初始化阶段，设置 a 为 1×10^{-3}。

用 srilm 工具[11]对来自于传统中医领域的 224 万平行句对的英文端数据(约含有 3600 万词语)训练 4 元语言模型，对来自于医药领域的 525 万平行句对的英文端数据(约含有 9800 万词语)训练 4 元语言模型。在翻译解码方面，用实验室内部研发的基于层次短语模型的解码器[9]作为机器翻译解码器，对开发集和测试集进行解码。该解码器使用的 8 个特征均来自文献[9]。

3.2 实验结果和分析

考虑到实验结果较多，本节先后给出在 batch EM 算法框架和 online EM 算法框架下，应用基于初始化的词语对齐增量式训练算法的实验结果。为描述方便，“传统中医”简称为 ctzy，“医药”简称为 yiyao。

3.2.1 Batch EM 算法下基于初始化的增量式训练

本小节将先后介绍传统中医领域和医药领域上增量式训练的实验结果。在 2.1 节中，针对模型的初始化策略细化出 4 种方案。这里针对这 4 种策略

① 具体包括以下 LDC 语料：LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2005T06, LDC2006E24, LDC2007E06, LDC2007E46, LDC2007E87, LDC2007E101, LDC2008E40, LDC2008E56, LDC2009E16 和 LDC2009E95。

② 开发集和测试集的句子是在之前提取的 224 万平行句对之外抽取出。对医药领域开发集和测试集的构建也是按照这个原则执行。

进行有关的对比实验。同时，把合并通用领域语料和特定领域语料后再训练的方法以及 1 种直接进行推断的方法也一同参与比较。为了描述方便，用下列符号来表示这几种策略。

1) Baseline: 直接在特定领域的语料上进行词语对齐模型训练，并最后生成词语对齐结果。

2) Com_out_in: 合并通用领域的语料和特定领域的语料，进行词语对齐模型训练。

3) Init_inference: 不在特定领域的语料上进行词语对齐模型训练，而是直接用通用领域语料上训练出的通用领域模型(包括词汇化翻译模型和扭曲模型)，在特定领域的语料上生成词语对齐结果。

4) Init_lex: 用通用领域语料上训练出的词汇化翻译模型的统计量做初始化，在特定领域的语料上进行模型训练。

5) Init_dist: 用通用领域语料上训练出的扭曲模型的统计量初始化，在特定领域的语料上进行模型训练。

6) Init_lex_dist: 用通用领域语料上训练出的词汇化翻译模型和扭曲模型的统计量初始化，在特定领域的语料上进行模型训练。

在应用层次短语模型解码器解码前，用开发集源语言端句子和测试集源语言端句子对抽取出的规则进行规则过滤，以加快解码的速度。表 1 给出传统中医领域上，各种方法经过源语言端句子过滤后的规则数目以及测试出的 BLEU 值。对比分析该表的若干数据，可以得到如下几个结论。

1) Init_inference 策略的翻译结果是最不好的，说明不在特定领域上进行训练而直接用通用领域的模型来生成词语对齐的方案是不合适的。

2) Com_out_in 策略的翻译结果是不好的，而且在测试集上的 BLEU 值比 baseline 的 BLEU 值低 1.10。说明，通用领域的语料和传统中医领域的语料是不同的，先简单地合并这两个语料得到一个整体语料，而后进行词语对齐的思路是不好的。我们认为原因是，语料的领域不同，使得在这个整体语料上的似然函数值最大(词语对齐的优化目标)并不能等同于在特定领域语料上的似然函数值最大。换句话说，在整体语料上通过模型的迭代训练得到的词语对齐模型参数，并不能很好地拟合特定领域的数据。

3) Init_dist 策略取得了最好的翻译结果，在测试集上的 BLEU 值比 baseline 的 BLEU 值高 1.92。与之形成对比的另两个方案，Init_lex 的 BLEU 值反而比 Baseline 的低 0.57，Init_lex_dist 的 BLEU 值比 Baseline 的高 1.21。从中可以看出，仅用词汇化翻译模型的统计量来做初始化，翻译效果反而下降。而用扭曲模型的统计量来初始化，翻译效果有显著提升，这可以从抽取出的规则数目较其他几种初始化方案抽取出的规则数目要至少多 37%的现象看出。我们认为产生这个现象的原因可能是，通用领域语料上的扭曲模型向特定领域语料上的模型训练提供了有效的指导，在这前提下，双语之间的词汇翻译即使不进行初始化，也能通过自行学习，并最终得到比 baseline 好的对齐结果。而仅仅用通用领域语料上的词汇化翻译模型来初始化，由于特定领域的数据量不大造成扭曲模型的训练得不到很好的指导，并最终导致了对齐结果的下降。

下面给出在医药领域的实验结果。实验配置上与传统中医领域的实验配置类似，不同之处仅在于平行语料、开发集和测试集更换为医药领域的数据。表 2 给出医药领域上，各种方法经过源语言端句子

表 1　词语对齐增量式训练方法在传统中医领域的对比结果(batch EM 算法)

Table 1　Performance of model training algorithms of word alignment in ctzy field (batch EM)

实验方案	开发集		测试集	
	规则数目	BLEU	规则数目	BLEU
Baseline	656730	45.78	767508	54.60
Com_out_in	1214593	44.18	1319314	53.50
Init_inference	350372	37.64	398289	45.26
Init_lex	552907	44.80	649959	54.03
Init_dist	**998083**	**46.93**	**1176400**	**56.52**
Init_lex_dist	724903	46.43	864063	55.81

表 2　词语对齐增量式训练方法在医药领域的对比结果(batch EM 算法)

Table 2　Performance of model training algorithms of word alignment in yiyao field (batch EM)

实验方案	开发集		测试集	
	规则数目	BLEU	规则数目	BLEU
Baseline	789259	18.73	659073	38.72
Com_out_in	1646098	19.08	1280424	38.64
Init_inference	113641	11.94	108331	26.58
Init_lex	728263	18.89	603377	38.52
Init_dist	1205741	19.14	1024683	40.74
Init_lex_dist	**1031885**	**19.27**	**862869**	**40.80**

过滤后的规则数目以及测试出的 BLEU 值。

总体上看，这些方案在医药领域上的表现与在传统中医领域上的表现类似。对实验结果的分析思路，也与对传统中医领域的分析思路类似。下面，简要总结下在医药领域上的几个重要的实验结果。

1) Com_out_in 策略的翻译结果仍然比 baseline 的翻译结果要差些，以测试集上的 BLEU 来看，Com_out_in 策略要比 baseline 策略差 0.08。

2) Init_dist 策略仍然是比 Init_lex 策略要好不少，以测试集上的 BLEU 值来看，其 BLEU 值比 Init_lex 策略高 2.22。不过，在这几种初始化策略中，Init_lex_dist 策略得到的翻译结果最好。

综合以上两组实验，可见 Init_dist 策略和 Init_lex_dist 策略是有效的基于初始化的增量式训练方法。

3.2.2 Online EM 算法下基于初始化的增量式训练

Online EM 算法有几个需要经验设置的参数——批量处理的规模 m 和步长衰减指数 α。这两个因素能够影响词语对齐模型训练的稳定性和收敛速度。考虑到这些，我们经验性地设置 m 为 1000，α 为 0.9。在这个设置下，为了得到 EM 算法迭代训练轮数的较好配置，首先在 2007 年全国统计机器翻译研讨会(SSMT 2007)中的词语对齐评测任务测试集(含有 504 句标注上词语对齐信息的双语句对)进行了一系列实验。以词语对齐常用的评价指标 AER 值和 F 值来看，发现当应用 online EM 算法时，IBM 模型 1 运行 3 轮，基于 HMM 的词语对齐模型运行 3 轮的模型训练迭代轮数的配置所得到的词语对齐质量，与应用 batch EM 算法时模型训练迭代轮数的默认配置所得到的词语对齐质量相当。换句话说，online EM 算法能有效提高模型的迭代收敛速度，减少迭代轮数，同时还不损失词语对齐质量。但至此还不能得到“不损失机器翻译质量”的结论。

这节通过一组实验来验证 online EM 算法在 IBM 模型 1 运行 3 轮，在基于 HMM 的词语对齐模型运行 3 轮的配置下所生成的词语对齐是否不损失机器翻译质量。语料的设置和上节的实验设置相同。由于实验结果看出，Init_lex_dist 方案表现既好又稳定，所以这里在 Init_lex_dist 的基础上，应用 online EM 算法替代之前用的 batch EM 算法，并进行有关实验(见表 3)。

将表 3 的传统中医领域的结果与表 1 进行对比，可以看出，同是基于初始化的增量式训练算法，以测试集上的 BLEU 值来看，online EM 算法的 BLEU 值比 batch EM 算法高 0.56。对于医药领域的结果，online EM 算法的 BLEU 值比 batch EM 算法仅提高了 0.02。需要强调的是，online EM 算法所用的迭代轮数是要少于 batch EM 算法的迭代轮数的。

表 3　词语对齐增量式训练方法在传统中医和医药领域的对比结果(Online EM 算法)

Table 3　Performance of model training algorithms of word alignment in ctzy and yiyao field (Online EM)

实验方案	开发集		测试集	
	规则数目	BLEU	规则数目	BLEU
Init_lex_dist (ctzy)	779276	46.68	920521	56.37
Init_lex_dist (yiyao)	1063634	18.96	894630	40.82

可见，在更少的迭代轮数上，online EM 算法能够得到与 batch EM 算法相当甚至是更好的机器翻译质量。所提出的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，既高效又保证词语对齐质量和机器翻译质量。

4　结语

统计机器翻译模型依赖于大规模双语语料库的训练。现有的训练方法在新增语料时需要合并新语料和原始语料，并重新进行模型训练。这种训练方法存在若干不足。本文围绕翻译模型构建流程的瓶颈——词语对齐，研究了既高效又能保证词语对齐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。实验表明：1) Init_dist 策略和 Init_lex_dist 策略是有效的基于初始化的增量式训练方法; 2) 在更少的迭代轮数上，online EM 算法能够得到与 batch EM 算法相当甚至是更好的机器翻译质量。可见，所提出的增量式训练方法既高效又保证词语对齐质量和机器翻译质量。该方法还能应用于其他基于无监督学习的自然语言处理问题，比如基于无监督学习的词性标注和句法分析。

为了更有效地解决翻译模型的增量式训练，还需要从短语/规则抽取及其分数计算方面来考虑。因此，下一步工作中将会考虑联合词语对齐和短语/规则抽取及其分数计算来研究解决翻译模型的增量式训练。

参考文献

[1] Wu Hua, Wang Haifeng, Liu Zhanyi. Alignment model adaptation for domain-specific word alignment // Proceeding of ACL 2005. Ann Arbor, USA, 2005: 467–474

[2] Duh K, Sudoh K, Iwata T, et al. Alignment inference and Bayesian adaptation for machine translation // Proceedings of MTSummit 2011. Xiamen, 2011: 114–121

[3] Brown P F, Stephen A. Della Pietra, Vincent J. Della Pietra, et al. The mathematics of statistical machine translation: Parameter Estimation. Computational Linguistics, 1993, 19(2): 263–311

[4] Vogel S, Ney H. Hmm-based word alignment in statistical translation // Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen, Denmark. 1996: 836–841

[5] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19–51

[6] Levenberg A, Callison-Burch C, Osborne M. Stream-based translation models for statistical machine translation // Proceeding of HLT/NAACL 2010. Los Angeles, 2010: 394–402

[7] Liang P, Klein D. Online EM for unsupervised models // Proceeding of HLT/NAACL 2009. Boulder, USA, 2009: 611–619

[8] Baum L E, Welch L R. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 1970, 41(1): 164–171

[9] Chiang D. A hierarchical phrase-based model for statistical machine translation // Proceeding of ACL 2005. Ann Arbor, USA, 2005: 263–270

[10] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of ACL 2002. Philadelphia, USA, 2002: 311–318

[11] Stolcke A. SRILM-an extensible language modeling toolkit // Proceedings of ICSLP 2002. Denver, USA, 2002: 901–904

[12] Koehn P, Och F J, Marcu D. Statistical phrase-based translation // Proceeding of HLT/NAACL 2003. Edmonton, Canada, 2003: 127–133