自纠正词对齐

龚慧敏 段湘煜 张 民
(苏州大学计算机科学与技术学院 苏州 215006)

摘 要 词对齐是统计机器翻译系统的重要一环,但词对齐的获得往往基于序列模型的计算,而没有考虑语言的结构化信息及语言特征,从而造成词对齐中出现一些不符合语言特征的结果。文中提出一种词对齐的自纠正机制,以纠正词对齐中的错误部分。该机制使用一些语言学上的先验知识,对词对齐结果进行由粗颗粒度到细颗粒度的纠正。首先采用基于标点的方法对句对进行粗粒度化纠正,然后采用基于统计特征的方法对子句对进行细粒度化纠正。该自纠正过程不需要借助任何其他词对齐工具和新语料。实验结果显示,自纠正词对齐显著提高了词对齐的准确率,并提高了机器翻译的质量,其中粗粒度的纠正方法对翻译质量的提高最为显著,细粒度的纠正方法也提升了翻译质量,最终通过结合粗颗粒度和细颗粒度的纠正方法,使翻译结果相对基准系统取得了显著的提高。

关键词 自纠正,词对齐,粗颗粒度到细颗粒度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.039

Self-correction of Word Alignments

GONG Hui-min DUAN Xiang-yu ZHANG Min
(Department of Computer Science & Technology, Soochow University, Suzhou 215006, China)

Abstract Word alignment is an important part of statistical machine translation systems. Previous works obtain word alignment through sequential models, which do not take into account the structure information and linguistic features of the language, leading to bad word alignments violating linguistic characteristics. This paper proposed a novel self-correction method for word alignments, aiming to correct the alignment errors which violate linguistic characteristics by exploiting linguistic prior knowledge. First, we conducted a coarse correction on short alignments obtained by binary segmentation based on punctuation method. Second, we proposed a fine-grained correction method for each short alignment based on statistical features. Third, corrected short alignments were merged to original alignments. This process does not rely on any third-party word aligner and additional parallel corpus. Experimental results show that our method significantly improves the accuracy machine translation results.

Keywords Self-correction, Word alignment, Coarse-to-fine

1 引言

统计机器翻译的核心思想是先对大量的双语平行语料进行统计分析,进而构建统计翻译模型,最终利用翻译模型对测试文本进行翻译。双语词对齐是统计机器翻译整体框架中非常重要的一个部分,它是短语表生成、调序规则抽取等的前提[1-4]。词对齐的准确率对统计机器翻译系统的性能有着不容忽视的影响。由于词汇的对译关系是翻译关系的基础,因此自动词对齐是跨语言信息处理的基础技术。

词对齐信息主要基于双语序列的统计信息[5-6],但并未考虑双语的层次结构和语言特征。如图1所示,依据图中的词对齐分布,英文中的两句对应中文的一句,中文以第一个逗号为界可分为两句,并与英文的两句一一对齐,但当中文句子分为两句后,会出现跨子句的词对齐问题。如实线框所示英文第二句的单词"to""the""in""the"等对齐中文第一句的"着""的","";而虚线框所示是子句中短语的词对齐越界问题,如英文短语"would like to"中的"would""to"对齐"我们","","elaborate on"中的'elaborate'对齐"我们""想"。上述词对齐错误是由于对所有可能的词对齐进行搜索,因此即使某些词对齐不符合语言特征也将被纳入词对齐的搜索空间中。在仅仅依靠双语序列的统计信息的条件下,即使这些错误词对齐不符合语言特征,也会因为较大的统计概率被输出。

文中提出了一种词对齐的自纠正机制,在传统词对齐的基础上引入了循环反馈,以上一轮的词对齐结果为输入,重新

到稿日期:2016-10-06 返修日期:2016-12-05 本文受国家自然科学基金:面向统计机器翻译的同步短语树结构规约机制研究(61273319)资助。

龚慧敏(1991-),女,硕士生,主要研究方向为机器翻译,E-mail:20144227036@stu.suda.edu.cn;段湘煜(1973-),男,博士,副教授,主要研究方向为自然语言处理、机器翻译;张 民(1970-),男,教授,博士生导师,主要研究方向为自然语言处理、机器翻译。

规划词对齐的搜索空间，以避免考虑不符合语言特征的词对齐，其自纠正词对齐结果如图 2 所示。传统词对齐中的虚线表示错误词对齐，在自纠正机制中，更新词对齐不仅避免了错误的词对齐，还增加了新的词对齐，如虚线标出部分。在该循环反馈的机制中，粗颗粒度到细颗粒度（coarse-to-fine）的方法被采用，句子级别的颗粒度被逐渐过渡到子句级别、短语级别的颗粒度来进行双语词对齐自动纠正。实验显示，首先采用基于标点的方法对句子进行粗颗粒度的切分，可以纠正跨子句的错误词对齐；然后在粗颗粒度化纠正的基础上，继而采用基于统计特征的方法对子句进行细颗粒度的切分，可以纠正短语之间的错误词对齐；最后，更新词对齐，其翻译结果相对基准系统取得了显著提高。可见，本文提出的自纠正词对齐方法显著提高了词对齐的准确率，并提高了机器翻译的质量。

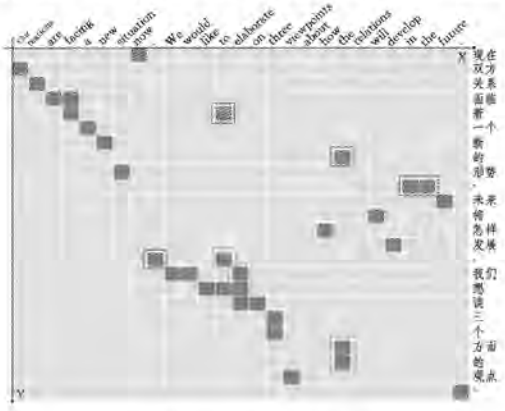

图 1　错误词对齐示例

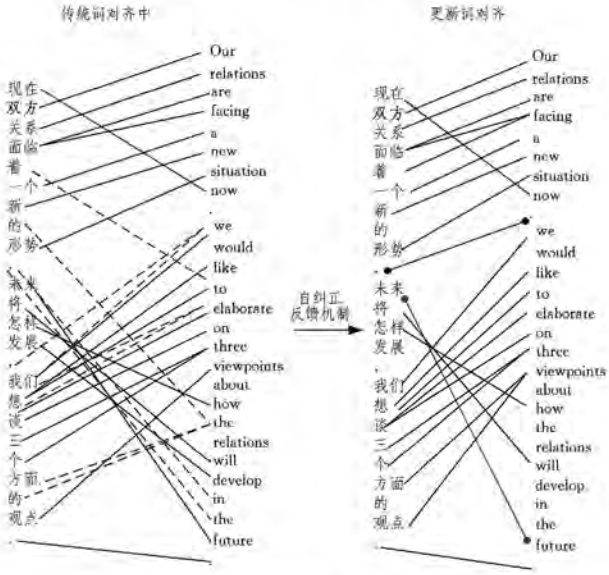

图 2　自纠正词对齐示例

2　自纠正词对齐

2.1　系统框架

自纠正词对齐不同于传统的词对齐架构，如图 3 所示，传统的词对齐模块位于虚线框之内，本文提出的自纠正词对齐在传统的词对齐模块的基础上引入了循环反馈环节，如图 3 中虚线框之外的部分所示。

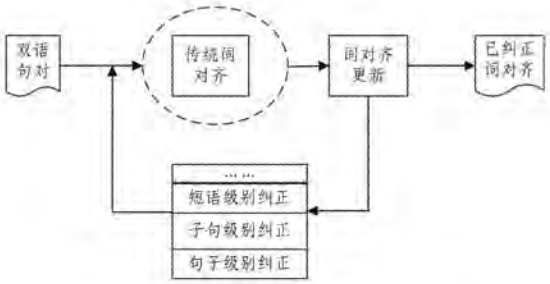

图 3　系统框架图

在该循环的反馈环节中，上一轮的词对齐结果通过虚线框之外的部分进行自纠正，自纠正过程包括：1）双语自纠正模块；2）词对齐更新模块。图 1 中汉英双语的句子不一致导致了图 1 中实线框部分的词对齐错误，为纠正此类错误，可将该双语句子提交至模块 1）进行句子级别的纠正，并将切分过的双语句对提交给传统的词对齐模块，这样可以保证词对齐结果一定符合双语句子级别的语言特征，避免对齐到双语句对之外的词对齐，同时也更正了在传统词对齐搜索空间的统计信息中错误的统计量。然后将传统词对齐模块的输出提交给模块 2）进行词对齐更新。上述过程循环反复来实现子句级别、短语级别的纠正，对上一轮的词对齐不断进行自纠正。

双语自纠正模块有助于剪去错误的词对齐搜索空间和更正词对齐的统计量。该模块主要依据两方面的语言特征：汉英双语句子级别的划分不一致的特征和汉英双语层次结构的语言特征。本文采用由粗颗粒度到细颗粒度（coarse-to-fine）的处理方法进行双语自纠正，首先针对句子级别进行切分，即粗颗粒度的双语切分；然后逐步针对子句级别、短语级别等层次进行细颗粒度的双语切分。具体说来，句子级别的纠正通过基于标点的方法完成，子句级别的纠正通过基于指示词（如从句引导词等）的方法完成，短语级别的纠正通过基于统计特征的方法完成。

词对齐更新模块与双语自纠正模块对应，由于被纠正的双语句对可能是原始句对的子句对，因此在合并已纠正的对齐子句时需要考虑原始双语句对的对齐情况。如原始双语句对是正序对齐时，子句从左到右进行合并；若是逆序对齐，则从右到左进行合并，从而更新词对齐。将已纠正的子句合拼成完整原句后，可以通过对比已纠正词对齐和传统词对齐来检验词对齐的更新情况。

在词对齐的自纠正系统框架中，双语纠正是至关重要的环节，以下将逐次介绍切分依据、基于标点、基于指示词和基于统计特征纠正的方法。

2.2　自纠正算法

源语言句子 S 由词序列 $s_1 s_2 \cdots s_n$ 组成，n 为 S 的长度；目标语言句子 T 由词序列 $t_1 t_2 \cdots t_m$ 组成，m 为 T 的长度；词对齐信息 A 由索引序列 $a_1 a_2 \cdots a_j$ 组成，其中对 a_j 的表示描述为：$a_j = j_s - j_t$，j_s 和 j_t 分别是 S 和 T 中相互对齐的词的索引值。如图 1 所示，英文句子排列在 X 轴，中文句子排列在 Y 轴，图中的所有方块代表词对齐信息。切分点 d 表示双语句对的划分处，本文采用坐标值形式对其位置进行描述，表示为

$d(s,t)$，其中 s 和 t 分别对应 S，T 中词的索引值，对应图 1 中横竖白色线条相交的点。

自纠正词对齐过程中，通过基于标点、基于指示词、基于统计特征等切分方法，实现了对原始双语句子从粗颗粒度到细颗粒度的切分来纠正词对齐，最终更新原始双语句对的词对齐。切分算法的伪代码如图 4 所示。

```
Step 1  句子级别纠正
  Step 1.1  标点组合列表 List_pc
  Step 1.2  for 句对 S,T:
  Step 1.3  依据 A,从 List_pc 中选出 d,将句对一分为二
  Step 1.4  更新词对齐纠正搜索空间,得到 A1 更新词对齐
Step 2  子句级别纠正
  Step 2.1  指示词字典 Dict_dw
  Step 2.2  for Step1 生成的句对:
  Step 2.3  依据 A1,结合 Dict_dw 确定 d,将句对一分为二
  Step 2.4  更新词对齐纠正搜索空间,得到 A2 更新词对齐
Step 3  短语级别纠正
  Step 3.1  for Step2 生成的句对:
  Step 3.2  依据 A2,依据统计特征确定 d,将句对一分为二
  Step 3.3  更新词对齐纠正搜索空间,得到 A3 更新词对齐
Step 4  进一步细颗粒度纠正搜索空间,直至搜索空间不再更新
```

图 4 自纠正算法伪代码

2.2.1 切分依据

(1)由于汉英两种语言构词及句法结构上的差异，在汉英双语句子中，句式结构可能是正序对齐，也可能是逆序对齐。因此，在对双语句子进行切分之前，需要判断句子的句式结构。本文通过统计词对齐的分布情况，即本文定义的词对齐密度 Φ 来判断正序逆序。词对齐密度 Φ 的计算公式如下：

$$\Phi=\frac{A_{s_b,s_e,t_b,t_e}}{|s_b-s_e|\times|t_b-t_e|} \tag{1}$$

其中，s_b，s_e，t_b，t_e 分别表示源语言端和目标语言端的起始和终止位置，这 4 个位置构成一个矩形区域，A_{s_b,s_e,t_b,t_e} 表示在该矩形区域中的词对齐个数。$|s_b-s_e|\times|t_b-t_e|$ 表示这个矩形区域的面积，Φ 是在这个矩形区域上的词对齐密度。

本文通过比较正序和逆序的词对齐密度的大小，来判断双语句对是正序对齐还是逆序对齐。当给定双语句对 S，T 的一个切分点 $d(s,t)$，$s\in(0,n)$，$t\in(0,m)$，令正序和逆序的词对齐密度分为 Φ_1 和 Φ_2，其计算公式如下：

$$\Phi_1=\frac{A_{0,s,0,t}}{s\times t}+\frac{A_{s,n,t,m}}{(n-s)\times(m-t)}$$

$$\Phi_2=\frac{A_{0,s,t,m}}{s\times(m-t)}+\frac{A_{s,n,0,t}}{(n-s)\times t}$$

若 Φ_1 大于 Φ_2，则句式结构为正序；否则为逆序。

(2)在切分过程中，为了从切分点集合中选择最佳切分点，本文定义了错误词对齐比值 e，并将其作为权衡切分点好坏的依据。在确定句式结构是正序还是逆序之后，根据切分点 d 统计偏离相应矩形区域的词对齐个数占所有词对齐的比例，即得到 e 值，计算公式如下：

$$e=\frac{A^{-}_{s_b,s_e,t_b,t_e}}{A} \tag{2}$$

其中，$A^{-}_{s_b,s_e,t_b,t_e}$ 表示不在矩形区域中的词对齐个数，A 为双语句对总词对齐个数，即 A_{s_b,s_e,t_b,t_e} 与 $A^{-}_{s_b,s_e,t_b,t_e}$ 相加之和。e 越小，其对应的切分点质量就越好。假设当前句式结构为正序，则切分点 d 对应的错误词对齐比值 e_z 的计算公式如下：

$$e_z=\frac{A^{-}_{0,s,0,t}+A^{-}_{s,n,t,m}}{A}$$

同理，逆序情况下的错误词对齐比值 e_n 的计算公式如下：

$$e_n=\frac{A^{-}_{0,s,t,m}+A^{-}_{s,n,0,t}}{A}$$

因此，在确定句式结构后，选取错误词对齐比值最小的切分点为最佳切分点。

2.2.2 基于标点的纠正

基于标点的纠正方法在本文多层次纠正模块中属于句子级别的粗颗粒度切分，将原始双语句对切分成两部分[7]。在基于标点的纠正方法中，标点包含英文的句号、分号、冒号和中文的逗号、分号、冒号，如表 1 所列。由于英文的句子单位往往比汉语的句子单位小，因此英文的逗号不在标点切分列表之内。穷举出双语句对中所有标点切分点的组合，为组合中的每一种情况计算相应的 Φ 和 e，在 Φ 一定的情况下，选择 e 最小的组合切分点作为最佳切分点，将双语句对切分为两组子句对。其切分过程具体如下：

1)读入的一组源、目标句子 S 和 T 以及对应词对齐 A；

2)检索句对 S 和 T 的标点，穷举出所有标点组合切分点；

3)按每个组合切分点将平面坐标切分成 4 个矩形区域，计算对应的 Φ 值；

4)基于 Φ 值寻找最小的 e 值所对应的切分点，对句对 S 和 T 进行切分，生成两组子句对。

表 1 中、英文标点

中文标点	、	；	：	
英文标点	.	;	:	

2.2.3 基于指示词的纠正

基于指示词的纠正方法是在上一轮的切分基础上对双语子句对进行切分，属于子句级别的细颗粒度的纠正。一般的双语句对中，英文句子会出现较多指示词，如关系代词、关系副词、连词等，与中文句中的代词、关系连词等相互对应，如表 2 所列。因此，根据中、英文指示词的相关特性，将双语句对一分为二。关于基于指示词的切分方法的相关过程类似于基于标点切分方法的切分过程，此处省略。

表 2 部分中、英文指示词

英文指示词	who, whose, that, which, and, or, but, as, why, thus, however, when, while, if …
中文指示词	她的，它的，他们的，和，或，但是，虽然，而且，既然，假如…

2.2.4 基于统计特征的纠正

基于统计特征的纠正方法是在短语级别中对双语句对进行细颗粒度的切分，即对句对进行二分。这里的双语句对是基于 2.2.2 节中的方法得到的子句对。针对双语句对的每个候选切分点构成特征向量 $X=\{x^1,x^2,\cdots,x^n\}$，计算其切分概率 $P(split=\text{true}|X)$，$split=\text{true}$ 表示进行切分。根据贝叶斯原理：

$$P(split=\text{true}|X)\propto P(split=\text{true})*P(X|split=\text{true}) \tag{3}$$

其中，$P(split=\text{true})$是一个均匀分布函数，表示双语句对中每个切分点为最佳切分点的概率相同；$P(X|split=\text{true})$是一个先验概率，表示最佳切分点为真的情况下统计特征向量 X 出现的概率，当集合中所有特征条件独立时，式(4)成立：

$$P(X|split=\text{true})=P(x^1x^2\cdots x^n|split=\text{true})$$

$$=\prod_{i=1}^{n}P(x^i|split=\text{true}) \tag{4}$$

其中，x^i 表示第 i 种特征，且本文假设变量 $x^i|split=\text{true}\sim Dirichlet(\tau)$，则其概率为：

$$P(x^i|split=\text{true})=\frac{Count(x^i,\text{true})+\tau}{Z} \tag{5}$$

其中，τ 是 Dirichlet 分布[8]的参数因子，表示特征出现的先验次数；$Count(x^i,\text{true})$表示在当前切分点是最佳切分点为真的情况下，特征 x^i 出现的频率；Z 是特征集合的总频率，是归一化因子。

由于在双语句对中没有对最佳划分点进行人工标注，因此采用 Gibbs 抽样算法对划分点的概率进行抽样逼近。当取得足够规模的样本后，这些样本将符合最佳划分点的真实分布，最终可在逼近的概率分布中选择概率最大的作为最佳切分点，Gibbs 抽样算法的伪代码如图 5 所示。

```
Step 1  初始化 Count(x^i,true)
  Step 1.1   for 每个句对：
  Step 1.2     初始化切分点列表 SP
  Step 1.3     for 每个可能的切分点 d：
  Step 1.4       SP.append(d)
  Step 1.5     从 SP 中随机取一个切分点为初始化切分点 d_init
  Step 1.6     获取 d_init 的特征向量 X
  Step 1.7   for x^i in X：
  Step 1.8       Count(x^i,true)++
Step 2  Gibbs 抽样
  Step 2.1   for 每轮抽样：
  Step 2.2     for 每个句对：
  Step 2.3       删除当前句的切分点的特征，Count(x^i,true)--
  Step 2.4     for 每个可能的切分点 d：
  Step 2.5       计算 d 的 P(x^i|split=true)，得到 P(X|split=true)
  Step 2.6     依据 P(X|split=true)，从 SP 中抽样出新的切分点 d_new
  Step 2.7     添加 d_new 的特征 x^i，Count(x^i,true)++
```

图 5　Gibbs 抽样算法的伪代码

本文中每个切分点的特征向量 X 包含 8 类特征，即 x^1-x^8，其特征模板为：切分位置的左右词特征及左右句法树结构特征，分别为 x^1-x^4 和 x^5-x^8。图 6 展示了双语句对 S 和 T 的部分句法树结构，椭圆圈表示切分点的特征，方框则是双语句对 S 和 T 的切分位置。对于 S 而言，其切分位置的左右词特征为“至此”“至今”，分别用 x^3 和 x^4 表示，其对应的句法树结构特征为 x^7 和 x^8，均为 VP。获取句法树结构特征 x^7 和 x^8 的过程为：将词汇特征 x^3 作为叶子节点向上遍历其句法树结构，直到发现某个节点与切分位置处于同一层，停止遍历，获取该节点的短语类型为特征 x^7；然后，依据词汇特征 x^4 同理获得特征 x^8。对于 T 而言，其过程类似，左右词特征为“Dynasty”“which”，分别用 x^1，x^2 表示，接下来，以 x^1 为叶子节点，向上检索句法树，直至找到某个节点和切分位置有一个共同的父节点，获取该节点的 NP 值作为特征 x^5；然后，以 x^2 为叶子节点，同理获取 SBAR 为特征 x^6。

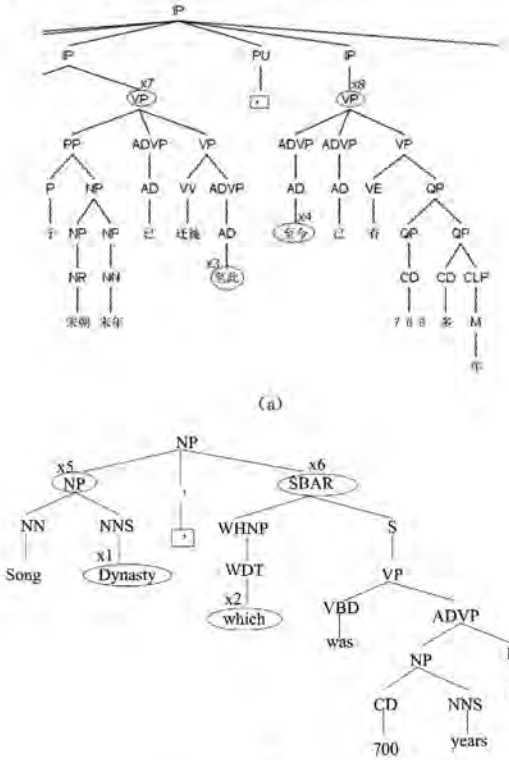

图 6　双语句对 S 和 T 的部分句法树结构及特征示例

3　实验分析

3.1　实验设置

本实验训练集所用汉英双语语料为 FBIS，包含 239416 条汉英句对；开发集为 NIST02 汉英双语语料，包含 877 条汉英句对，测试集包含 NIST03－NIST05 等 3 个汉英双语语料，其中每个源句对应 4 个目标译文。使用 MOSES 训练基统计机器翻译模型。首先，汉到英方向和英到汉方向分别运行 MGIZA++[9]工具，采用“grow-diag-final”的启发式方法获得词对齐信息；其次，使用 SRILM[10]工具在新华英文语料集 Gigaword 上训练五元语言模型；然后，通过 MERT 调整最小错误率训练过程中各个特征的权重值[11]；最后，采用 BLEU 评测标准评价翻译质量。

在实验数据的准备过程中，汉英双语语料的分词分别采用中科院中文分词工具、MOSES 的 token 函数。然后，对双语句对进行过滤：对于目标句子译文并不是源句子的翻译的情况，以及源句子长度与目标句子长度比值超过正常匹配阈值的情况，均选择过滤。出现上述两种情况的双语句对基本是源句子与目标句子不匹配的问题所致，在无监督情况下是无法纠正的，属于双语语料的噪音，因此一旦识别出来本文将不处理，目的在于减小实验过程中的误差。

实验过程中，设定每个方法中的限制条件：1)基于标点切分的方法中，汉英双语句对排除句尾标点，句中必须均存在如表 1 所列的标点符号，才能对双语句对进行切分处理；2)基于指示词的切分方法中，双语句对的长度均大于 15 时，才进行切分处理；3)基于统计特征的切分方法中，式(5)的先验因子 τ 值设为 1，以平衡统计特征出现的次数，避免当特征出现次数为 0 时，概率为 0 的情况出现。

3.2 实验结果分析

3.2.1 词对齐纠正分析

实验中为了避免实验结果的局部偏差，评测时均采用 3 次抽样结果的均值作为统计结果，每次抽样 200 句。实验分析使用的评测标准包括句子切分准确率 ACC 和词对齐错误率 AER[12](Alignment Error Rate)。

为了评价每种纠正方法的准确率，本文通过统计样本中被正确纠正的句子数 a 占需要被纠正的句子数 b 的比例来计算 ACC，即 $ACC=a/b$。

AER 是在分析词对齐错误的过程中经常用到的评价标准，其定义如下：

$$AER(S,P,A)=1-(|A\cap S|+|A\cap P|)/(|A|+|S|) \tag{6}$$

其中，S 表示正确的词对齐集合，P 表示可能正确的词对齐集合，A 则是本文方法输出的词对齐集合。然而，本文在分析词对齐纠正的过程中，对 P 和 S 不加以区分，均为正确的词对齐 S，故本文中对 AER 的定义如下：

$$AER=1-\frac{2\times|A\cap S|}{|A|+|S|} \tag{7}$$

如表 3 所列，SBP，SBK 和 SBG 分别表示基于标点、基于指示词和基于统计特征的纠正方法，AER_{giza}，AER_{self} 分别表示传统词对齐工具 GIZA 的词对齐错误率和进行自纠正后的词对齐错误率。在实验数据的统计过程中，通过 ACC 值可知，在 SBP 方法中，原始双语句对约有 29.2%的句对被纠正，句子纠正准确率分别高达 95.6%；在 SBP 更新的数据基础上，SBK 与 SBG 方法中，约 78.6%的句对被纠正，句子纠正准确率分别为 72.0%和 93.3%，显然，基于指示词的纠正方法的准确率明显低于基于统计特征的纠正方法。

表 3 每种方法的 ACC 值和 AER 值

纠正方法	ACC	AER_{giza}	AER_{self}
SBP	0.956	0.162	0.093
SBK	0.720	0.093	0.097
SBG	0.933	0.093	0.079

上述结果主要原因在于基于指示词的纠正方法在子句级别的纠正过程中存在两点缺陷：1)指示词字典由于构建不全面，导致对最佳切分点的确定存在一定的偏差，从而造成切分错误；2)盲目检索指示词，造成指示词滥用，如英文的指示词“and”，当其充当两个事物或人物名词的连接词时，相对于整个句子，它并不是一个好的切分点，因此，如果将同样具有该性质的指示词作为切分点，会造成子句间严重的切分错误。然而，在基于统计特征的纠正方法中，对基于指示词的方法做出了改进，为每个切分点构建了特征属性：汉英文左右词特征和句法树结构特征。因此，可计算出每个切分点的切分概率 $P(X|split=true)$，进而选择概率最好的作为切分点。因此，在基于统计特征的纠正方法中，如上述的错误切分点的概率均偏低，可通过句法结构特征识别出来，从而避免错误纠正，这也是基于统计特征的纠正方法优于基于关键词的纠正方法的关键。

比较每种方法下的 AER_{giza} 和 AER_{self} 值发现，基于标点的纠正方法对错误词对齐的纠正效果显著，可见句子级别的纠正简单有效；然而，在完成句子级别纠正的基础上，基于指示词的纠正方法的效果并不理想，其 AER 并没有得到有效的改善。因此，本文在接下来的细颗粒度纠正过程中，采用了改进的方法——基于统计特征的纠正方法，使得 AER 值进一步降低，并证明了该方法能有效地纠正错误词对齐。

3.2.2 翻译结果分析

本文的基准系统 Baseline 是 GIZA++的对齐结果；SBP 为在 Baseline 的基础上按基于标点的纠正得到的词对齐结果；SBP+SBG 表示在 SBP 数据上按基于统计特征的纠正得到的词对齐结果；$Update_{SBP}$ 是更新 SBP 的结果后得到词对齐的整体纠正结果；同理，$Update_{SBP+SBG}$ 是更新 SBP+SBG 的结果后得到的整体词对齐纠正结果；Cat_{F1} 表示将 Baseline 与 $Update_{SBP}$ 的数据叠加后得到的数据；同理，Cat_{F2} 是叠加 Baseline 与 $Update_{SBP+SBG}$ 得到的数据。在上述各个方法获得的词对齐数据的基础上进行短语模型的训练和在测试集上的解码，其在不同测试集上的性能如表 4 所列。

表 4 不同 NIST 测试集合的 BLEU 值

SYSTEM	NIST03	NIST04	NIST05
Baseline	27.37	27.46	27.53
SBP	28.08	28.10	28.14
$Update_{SBP}$	28.13	28.15	28.19
Cat_{F1}	28.56	28.63	28.62
SBP+SBG	28.21	28.33	28.37
$Update_{SBP+SBG}$	28.40	28.43	28.48
Cat_{F2}	28.78	28.81	28.86

通过比较 BLEU 值可知，基于标点纠正的数据 SBP 的 BLEU 值比 Baseline 提高了 0.66，主要是由于其对双语句对按标点组合进行切分，纠正了跨句子的词对齐，从而提高了词对齐的对齐质量及翻译质量。$Update_{SBP}$ 的翻译结果与 Baseline 相比，提高了 0.71，主要原因在于在合并切分的子句时，其词对齐在被纠正的情况下同时合并，纠正的词对齐相对于未纠正的词对齐的对齐情况更好。Cat_{F1} 的数据包含了传统词对齐和已纠正的词对齐，其中已纠正的词对齐剔除了不匹配对齐情况，并纠正了跨句子间的词对齐问题，与传统的词对齐相辅相成，统计信息亦相互补充，与 Baseline 相比提高显著，为 1.15。

SBP+SBG 是在基于标点纠正的数据上，进一步按基于统计特征纠正的细粒度切分后得到的数据，其翻译结果相比 SBP 有所提升，因为基于统计特征的纠正对句子进行了更细颗粒度的切分，考虑了切分点左右词及句法树结构，依据这些统计特征信息确定切分位置，可以纠正短语级别的词对齐问题；与 Baseline 翻译结果相比，有效提高了 0.85。$Update_{SBP+SBG}$ 与 $Update_{SBP}$ 相比，BLEU 值得到了进一步改善，

(下转第 238 页)

analysis using Big Data technology[C]// International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE). 2015:467-471.

[6] DAM R V D. Big Data a Sure Thing for Telecommunications: Telecom's Future in Big Data [C]// Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). 2013:148-154.

[7] OUYANG Y, HU M M, HUET A, et al. Mining of leaders in mobile telecom social networks [C]// Wireless Telecommunications Symposium (WTS). 2016:1-4.

[8] HUANG W L, CHEN Z, DONG W Y, et al. Mobile Internet big data platform in China Unicom [J]. Tsinghua Science and Technology, 2014, 19(1):95-101.

[9] CHETAN S B J, SRINIVASA K G. Large Scale Multi-label Text Classification of a Hierarchical Dataset using Rocchio algorithm [C]// International Conference on Computational Systems and Information Systems for Sustainable Solutions. 2016:291-296.

[10] YANG W C, FU Y M, ZHANG D. An Improved Parallel Algorithm for Text Categorization [C]// International Symposium on Computer, Consumer and Control. 2016:451-454.

[11] SANTOSO J, YUNIARNO E M, HARIADI M. Large Scale Text Classification using Map Reduce and Naïve Bayes Algorithm for Domain Specified Ontology Building [C]// 7th International Conference on Intelligent Human-Machine Systems and Cybernetics. 2015:428-432.

[12] YANG J, YANG M H. Top-Down Visual Saliency via Joint CRF and Dictionary Learning [C]// Computer Vision and Pattern Recognition. IEEE, 2012:2296-2303.

(上接第 220 页)

可见由粗颗粒度到细颗粒度的多层次纠正，可纠正子句间、短语间的错误词对齐；与 Baseline 相比，实验结果提高了 0.99，证明了本文提出的自纠正词对齐方法能有效提高词对齐质量和机器翻译的质量。Cat_{P2} 的 BLEU 值与 Cat_{P1} 相比，依旧有所改善；与 Baseline 相比提高了 1.37，提升效果显著。

结束语 本文提出了一个针对词对齐的自纠正机制，借助于语言特征等先验知识，对词对齐进行多轮循环自纠正。在粗颗粒度到细颗粒度的纠正过程中，首先在粗颗粒度的级别上采用基于标点的纠正方法，对原始双语句对进行句子级别的切分，该方法保证了句法结构的完整，方法简单有效，准确率高达 95.6%，AER 结果改善明显；然后在子句颗粒度的级别上采用基于指示词的纠正方法，对上述切分的子句进行细颗粒度切分，发现准确率偏低，为 72.0%，而且 AER 结果并没有得到改善；另外，在短语级别的颗粒度上采用基于统计特征的纠正方法，对上一轮切分的子句进行细颗粒度切分，准确率高达 93.3%，AER 结果较基于标点的切分方法进一步得到改善。通过分析发现，在细颗粒度的词对齐纠正中，基于统计特征的切分效果明显优于基于指示词的切分，因此，本文在自动纠正词对齐过程中，先采用基于标点的方法、后采用基于统计特征的方法对双语句对进行纠正，最后将切分的子句合并成完整句子，翻译质量得到了显著提升。

参 考 文 献

[1] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003:127-133.

[2] LIU Y, LIU Q, LIN S. Tree-to-string alignment template for statistical machine translation[C]// International Conference on Computational Linguistics and, Meeting of the Association for Computational Linguistics(ACL 2006). Sydney, 2006:609-616.

[3] GALLEY M, GRAEHL J, KNIGH K, et al. Scalable inference and training of context-rich syntactic translation models[C]// International Conference on Computational Linguistics and the, Meeting of the Association for Computational Linguistics. 2012:961-968.

[4] CHIANG D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2):201-228.

[5] BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2):263-311.

[6] LIANG P, TASKAR B, KLEIN D. Alignment by agreement [C]// North American Association for Computational Linguistics (NAACL). 2006.

[7] XU J, ZENS R, NEY H. Partitioning parallel documents using binary segmentation[C]// The Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2006:78-85.

[8] BLUNSOM P, COHN T, GOLDWATER S, et al. A Note on the Implementation of Hierarchical Dirichlet Processes[C]// International Joint Conference on Natural Language Processing of the Afnlp. DBLP, 2009:337-340.

[9] GAO Q, VOGEL S. Parallel implementations of word alignment tool[C]// Association for Computational Linguistics. 2008:49-57.

[10] STOLCKE A. SRILM-an extensible language modeling toolkit [C]// Proceedings of the 7th International Conference on Spoken Language Processing. 2002:901-905.

[11] OCH F J, NEY H. A systematic comparison of various statistical alignment models [J]. Computational Linguistics, 2003, 29(1):19-51.

[12] OCH F J. Minimum error rate training in statistical machine translation [C]// Meeting on Association for Computational Liguistics. 1973:160-167.