双语语料库段落重组对齐方法研究

李维刚 刘挺 王震 李生
哈尔滨工业大学计算机学院 信息检索研究室 哈尔滨 150001
E-mail:{lee, tliu, wangzhen, ls}@ir.hit.edu.cn

摘要：网络上存在的大量双语资源，给构建大规模双语语料库提供了可能。双语对齐作为语料库加工过程中的关键技术，已经引起研究者的高度重视。针对目前可收集到的双语资源大都没有做到段落对齐，本文结合基于句子长度和基于词典的两种经典对齐算法思想，充分利用双语文本中的句子在整个文本中的位置信息，在(1:1)型句珠里选取锚点，并根据双语文本特征引入一部双语词典进行校验，从而获得分段的锚点，实现通用的段落重组对齐。

关键词：双语语料库，段落重组对齐，锚点，匹配

Research of Paragraph Realignment of Bilingual Corpus

Li Weigang Liu Ting Wang zhen Li Sheng
Information Retrieval Laboratory Harbin Institute of Technology, Harbin 150001
(bert.tliu.zsf.car.qinb.ls@ir.hit.edu.cn)

Abstract: Large amount of bilingual resource on the internet bring the probability of building a large scale of bilingual corpus. As the key technology during the course of building the corpus, bilingual alignment technology is growing high recognition. Facing the situation that most of bilingual resource attained on the internet is aligned in paragraph, paragraph realignment is necessary. Combining length-based method with lexicon-based method, making full use of the location information of each sentence in whole text, we choose the anchors among the 1-for-1 beads according the result of dictionary check and achieve the goal of general paragraph realignment.

Keywords: bilingual corpus, paragraph realignment, anchor, match

1. 引言

双语语料库是一种包含有两种语言互译信息的特殊的语料库。它能够提供两种语言之间丰富的匹配信息，在翻译知识的获取、双语词典的建立、基于实例的机器翻译、词义消歧等领域有着重要的应用价值[1]。

大规模双语语料库的建设是进行基于语料库研究的基础，它包括语料库的设计、语料的

采集、录入和管理等方面[2]。而目前互联网上存在着丰富的双语资源，为短期内建成大规模的双语语料库提供了可能。因此对网上可收集到的双语互译文本进行加工成为一个非常有意义的课题。

对齐技术是加工双语文本的核心。所谓对齐就是从互译的不同语言文本中找出互译片断的过程，双语语料库对齐可分为段落、句子、短语、单词等不同级别的加工深度 ，语料库的加工深度决定了语料库所能提供的知识的粒度。早在 90 年代初期，国外就有人开始这一方面的工作，主要有 Brown[3], Gale[4] 和 Chen[5]等，他们的方法主要归结为两类，基于长度的对齐方法和基于词汇的对齐方法，Brown 在对 Hansard 语料库进行对齐时，引入了锚点(anchor)的概念，认为锚点的作用就是将整个语料库分成一些小的对齐片断；同时把每一对相对应的句子称作句珠(Sentence bead)。针对汉英双语对齐，国内的刘昕[6]，钱丽萍[7]等人也进行了一些改进的对齐算法研究。目前很多学者在进行双语对齐研究时，大多数都是在段落对齐的基础上进行句子对齐的研究。然而目前网络上的大量双语文本基本都没有做到段落对齐，而段落对齐是进行后续的句子对齐、结构对齐等更深级别的对齐加工的基础，因此针对这种真实的文本资源，必须首先进行段落对齐。

文献[1]提出了一个将文本依照翻译块(translation block)重新进行分段的方法，它通过汉英词汇对之间的特征比较，首先对汉语句子进行分词，找到可以用于汉英语料库分段的锚点词汇对，在此基础上，通过锚点词所在句子的匹配获得锚点句子对来进行分段。但是这种方法仅适合于具有较多高频固定词的双语文本的分段对齐，对于只具有较少高频固定词的双语文本，这种方法就会遇到数据稀疏问题导致分段太粗及准确率下降。通过对网络上收集到的数百篇、各种体裁的真实文本考察，发现 90%以上的电子文本中的段落并不对应或者没有明显的段落标记，如果进行自然段的对齐实现比较困难，并且分段太粗，因此针对这种情况有必要进行重新分段。本文提出将两篇互译的双语文本各看成一个整体，对文本中段落进行重新组合后对齐。充分利用双语文本中的每一句在整个文本中的位置信息，以(1:1)型句珠作为候选锚点，并根据双语文本特征引入一部双语词典进行校验，从而获得分段的锚点句，实现通用的段落对齐。

2. 段落对齐的问题描述

对一个双语平行文本的段落对齐，就是要找出两种语言文本中段落之间的对应关系，那么对齐后的文本就表现为具有相等段落总数的互译组块序列。

本文通过扩充图论中二分图及匹配的概念，给出段对齐的形式化描述：

二分图(Bipartite Graph) 设 G 为无向图，G=<V，E>，结点集 V 有两个子集 V_1，V_2 满足 $V_1 \cup V_2=V$，$V_1 \cap V_2=\phi$，使 G 的每一条边 $e \in E$ 时，$e=\{v_i,v_j\}$，$v_i \in V_1, v_j \in V_2$，即同一子集 $V_i(i=1, 2)$中的任何两个结点都不邻接，称这样的图为二分图。G 记为 $G=<V_1, E, V_2>$。对于二分图 $G=<V_1, E, V_2>$中，若 V_1 的每个结点与 V_2 的每个结点相邻接，反之亦然。则称 G 为完全二分图，若$|V_1|=m$，$|V_2|=n$，则简记为 $K_{m,n}$，如图 1 所示。

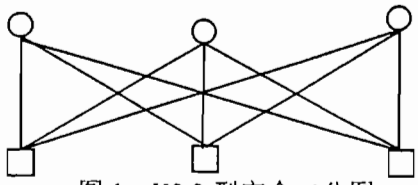


图 1　K3,3 型完全二分图

匹配(Matching)　设 $G=<V_1,E,V_2>$是二分图，如 E 的一个子集 M 中的边无公共端点，即任两边均不邻接，则 M 为 G 的一个匹配。

本文所研究的段落重组对齐，可以认为是一个特殊的匹配问题，为此我们定义这种模型为“最优对齐匹配”。

M 为完全二分图 $G=<S,E,T>$的一个最优对齐匹配，如果二分图中所有的节点是有序的，并且 M 中存在的任意一条边 $e=\{s_i,t_j\}$的权值为 $d(s_i,t_j)$，必须满足 $d(s_i,t_j)< D$（D 为特定的阈值）；此时，M 中不存在边$\{s_k,t_r\}$使得 $k<i$ 且 $r>j$ 或 $k>i$ 且 $r<j$ 成立；若$|S|=m$，$|T|=n$ 则首先默认$\{s_m,t_n\}\in E$，在满足上述条件前提下，依次在完全二分图 G 中选取权值最小的边，直到 M 中边数达到饱和。

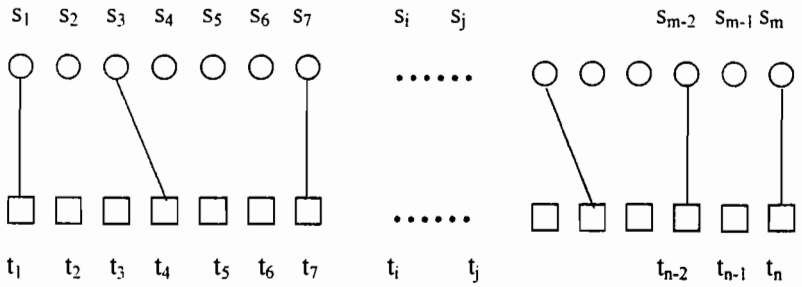
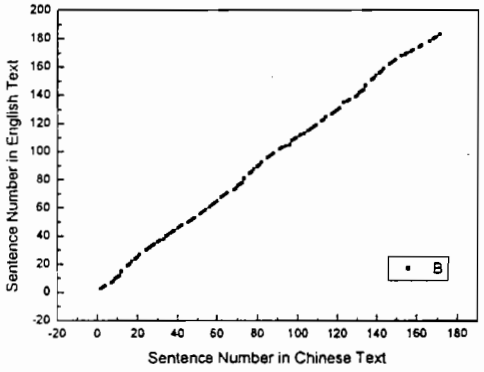


图 2 在阈值 D 限定下的一个 Kn,n 最优对齐匹配示意图

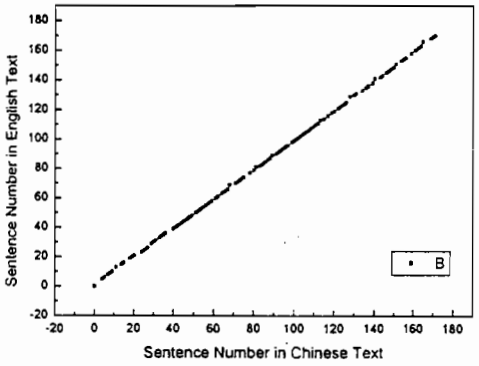
文本 S 中每一句子对应 V_1 中一个顶点，文本 T 中每一句子对应 V_2 中一个顶点，则双语文本可以分别表示为 $S(s_1, s_2, s_3,\ldots s_i, \ldots s_j, \ldots s_m)$和 $T(t_1, t_2, t_3\ldots t_i, \ldots t_j, \ldots t_n)$，不考虑语义仅从形式上讲，S 和 T 中任何两个元素 s_i 和 t_j 均可作为一个(1:1)型句珠(s_i,t_j)的两个元素，构成一个完全二分图 $K_{m,n}$。因此双语段落重组对齐的过程就是在一个完全二分图中寻找一个“最优对齐匹配”的过程。图 2 中存在边 $e=\{s_i,t_j\}$，我们则视为文本 S 中的第 i 句和文本 T 中的第 j 句对应，为了达到分段效果，我们在 S 中的第 i 句和文本 T 中的第 j 句后分别做一“段标”。为了使对应文本 S 和 T 正常结束时，S 和 T 中有一样的段落数，我们默认文本 S 和 T 中的最后一句是对应的，即在数学模型最优对齐匹配中，若$|S|=m$，$|T|=n$ 则首先默认$\{s_m,t_n\}\in E$ 成立。最优对齐匹配的概念保证了选取的句对中不存在一对多或交叉对应的情况，依次在完全二分图 G 中选取权值最小的边，也就意味着每次选出来的句对，均是在未对齐的句子中最有可能构成(1:1)型句珠的句对。M 中边数达到饱和，即在特定条件下（一定的 D 值），双语段落重组对齐完成。

3．段落对齐的锚点选择

我们考察了(1:1)型句珠在各种体裁、不同风格的文本中的分布情况，将人工对齐结果中(1:1)型的句珠选取出来，横坐标为句珠的中文句子在其所属的文本中编号，纵坐标为句珠的英文句子在其所属的文本中编号。几乎所有的显示结果都如图 3 所示：



(a)双语文本句子总数相差较大



(b) 双语文本句子总数相差较小

图 3 双语文本中(1:1)型句珠分布情况

3（a）为互译的英汉文本中句子总数相差较大时的分布情况，3（b）为句子总数相差较小时的分布情况。经大量统计发现句珠类型为(1:1)的句对在全篇的比例均超过 85%，并且有着良好的分布规律。文献[8]中也提到，经过大量统计，中英文双语对齐后，句珠类型为(1:1)的句对在全篇的比例为 89%。如果我们可以选择这种(1:1)型的句珠作为候选锚点，将使得段落对齐具有通用性。因此我们提出一种新方法，其总体思想可以表述为：篇章定位，句长搭配，词典校对。即为利用对应句子在文本中的位置信息，对每组句子均考虑其在全篇中的位置，利用句子前面的段落和句子后面的段落及对应句子本身的长度，计算出完全二分图中的每一个边的权值，再经过词典校验选出锚点句对。

考虑互译文本 S 和 T，他们的总长度分别表示为 Ls 和 Lt，文本 S 中第 i 句和文本 T 中第 j 句，分别用 s_i 和 t_j 表示，s_i 上部文本总长度为 Us_i；本句长度为 Ls_i；下文部分语句总长度为 Ds_i；同样，t_j 上部文本总长度为 Ut_j；本句长度为 Lt_j；下部文本总长度 Dt_j；

考虑句子在上下文中的位置信息，研究 s_i 和 t_j 的关系时，定义四个比值：

对应文本长度之比：$P0 = Ls/Lt$ ；

对应上文部分长度之比：$Pu[i,j] = Us_i/Ut_j$ ；

对应句长度之比：$Pl[i,j] = Ls_i/Lt_j$ ；

对应下文部分长度之比：$Pd[i,j] = Ds_i/Dt_j$ ；

若 si 和 tj 确实可以构成(1:1)句珠时，则 Pl[i,j]将在一特定区间之内。我们假定对应句长度之比 Pl[i,j]在表明对应信息中加权系数为 1，仅从形式上考虑，对应上文长度之比 Pu[i,j]和对应下文长度之比 Pd[i,j]等价，因此在表明候选锚点对应信息时应具有相同的权重 a，可构造下面的一个形式对齐评价函数：

$$P[i,j] = a(Pu[i,j]-P0)^2 + (Pl[i,j]-P0)^2 + a(Pd[i,j]-P0)^2$$

函数中加权系数 a 意义：使上下文长度和本句的长度对评价函数值均起到相应的作用。在我们的试验中加权系数 a 由当 i=m/2，j=n/2 时的$(Pu[i,j]-P0)^2$值和形式对齐阈值 D 的大小共同调节而定。实验证明，由于采用贪心算法，最优对齐匹配过程中，所有数值的大小均是相

对的，所以加权系数 a 在一定的范围内即可得到良好的试验结果。从形式上考虑，评价函数值 P[i,j]越小，则 s_i 和 t_j 可以构成句珠的可能性越大。

仅从形式上选取锚点，无法保证准确率，为了确保段落对齐的准确性，我们在对选出的 P[i,j]校验之后，又进行了词典检验。根据文献[9]提出的公式如下：

$$H = \frac{L|Match(S)| + L|Match(T)|}{L|S| + L|T|}$$

其中，L| |表示全部元素的字节长度和；Match(S)代表（根据英汉词典）译文中出现在汉语句子中的英语单词，Match(T)代表（根据英汉词典）成为英文单词译文的汉语单词。符合词典校验的句珠将成为段落重组对齐的锚点。

4. 段落重组对齐算法实现

1) 读入英汉双语文本和英汉词典；
2) 识别英语和汉语句子边界，对每一个句子进行编号；
3) 在双语文本开始各虚拟添加一理想长度以平滑文本；
4) 默认双语文本中最后一句是对齐的，计算其它所有形式对齐评价函数值 P[i,j]；
5) 选出形式对齐评价函数值最小的 $P[i_1,j_1]$；
6) 如果 $P[i_1,j_1]<D$，则对选定的双语句对进行词典校验，否则转步骤 8)；
7) 如果 H>H0（句对词典校验阈值 H0），则将 P[i,j]中 $i>i_1$ 且 $j<j_1$ 和 $i<i_1$ 且 $j>j_1$ 的区域置为无效，分别在选中的中英文句尾加段标，否则将 $P[i_1,j_1]$ 置为无效；返回 5)；
8) 输出段对齐后的文本，结束。

5. 实验结果与分析

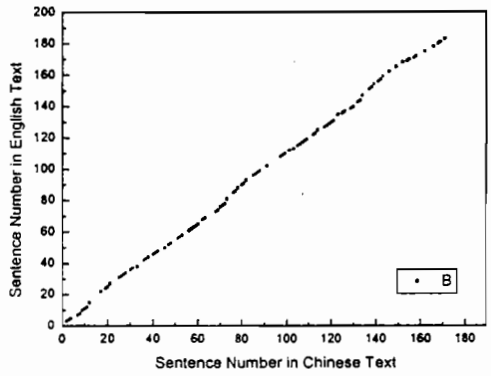
我们试验中所收集的上百兆双语篇章级对齐文本，包括各种文学作品、新闻以及政治、法律类双语材料，经统计 95%以上的篇幅大小在 500 句以内，对此我们使用算法的良好效果只是在处理单篇 500 句以内的文本时得到验证。

由于是重新分段，因此召回率的概念我们就定义为：

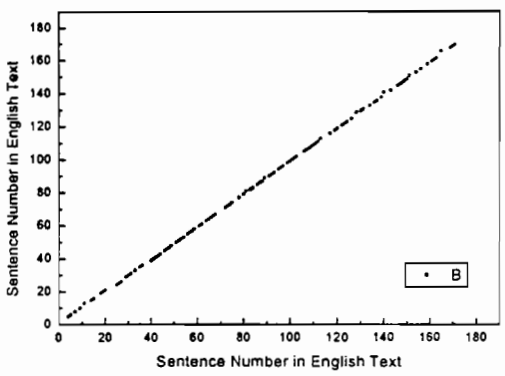
召回率 = （返回段落总数 – 错误段落总数）/ 双语文本中(1:1)型句珠总数；

准确率 = （返回段落总数 – 错误段落总数）/ 返回段落总数；

对应图 3 的双语文本，图 4 为相应的锚点分布图。



(a)双语文本句子总数相差较大



(b) 双语文本句子总数相差较小

图 4　双语文本中锚点提取情况

不难看出两者点的分布规律完全相同，只是点的稀疏程度不同。经过本算法段落对齐后，每段句子数 3-5 句，段落分布均匀，对于对译程度较好的文本，经过段落重组对齐后，在文本的某些部分甚至可达到句子级对齐效果。

表 1 对齐试验结果

	图 3(a)对应文本	图 3(b)对应文本
(1:1)型句珠总数	158	169
提取锚点总数	108	121
准确率	99.07%	100%
召回率	68.35%	71.60%

表 1 中对齐试验结果可以看出，这种利用(1:1)型句珠作为段落重组对齐锚点的方法，所依赖的形式对齐评价函数值是在全篇的基础上得出的，因此匹配过程中可以将错误控制在最小的范围内，与传统的基于长度的对齐方法相比，有效的抑制了错误蔓延。即使在处理有少量句子省略的文本时，该算法同样体现出很强的鲁棒性。在进行词典检验时，无需对汉语分词，实现简单，代价较小，能够在保证准确率的基础上，实现分段均匀并且段落很细的效果。

6. 结束语

本文提出一种新的段落重组方法，其主要思想是：充分利用句子在文本中的位置信息，对文本中大于 85%的(1:1)型句珠进行提取，然后采用一部英汉双语词典，从语义上保证了锚点选择的准确性，实现了段落重组对齐。由于选择了(1:1)型句珠作为候选锚点句，因此这种方法具有很强的通用性。我们下一步的工作将在段落重组对齐的基础上，将此算法引入到句子对齐和其它语种的对齐。

参 考 文 献

[1] 王斌. 刘群. 张祥. 汉英双语自动分段对齐研究. 软件学报, 2000, 11(11):1547-1553

[2] 黄昌宁. 李涓子. 语料库语言学. 北京：商务印书馆, 2002, 17-20

[3] Brown. P.F.. Cocke. J.. Della Pietra. S.A.. et al. A Statistical Approach to machine translation. Computational Linguistics. 1990. 16(2): 79-85

[4] Gale. W.A. Church. K.W. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics. 1993. 19(2): 75-102

[5] Chen. Stanley. Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL). 1993

[6] 刘听. 周明. 朱胜火. 黄昌宁. 基于自动抽取词汇信息的双语句子对齐. 计算机学报, 1998. 21(8): 151-158

[7] 钱丽萍. 赵铁军. 杨沐昀. 高光来. 基于译文的英汉双语句于自动对齐. 计算机工程与应用, 2000, (12): 59-61

[8] Dekai Wu. Aligning A Parallel English-Chinese Corpus Statistically with Lexical Criteria. Meeting of the Association for Computational Linguistics.1994

[9] 杨沐昀. 汉英句子对齐及翻译词典和翻译规则的自动获取. 哈尔滨工业大学博士论文. 2002