文章编号：1003-0077(2008)02-0087-05

基于自动句对齐的相似古文句子检索

郭 锐，宋继华，廖 敏

(北京师范大学 信息科学与技术学院 北京 100875)

摘 要：随着语料库语言学的兴起，基于实例的机器翻译(EBMT)得到越来越多的研究。如何快速准确地构建大规模古今汉语平行语料库，以及从大量的对齐实例(句子级)中检索和输入句子最相似的源句子是基于实例的古今汉语机器翻译必须解决的问题。本文综合考虑句子长度、汉字字形、标点符号三个因素提出了古今汉语句子互译模型，基于遗传算法、动态规划算法实现了古今汉语的自动句对齐。接着为古文句子建立全文索引，基于汉字的信息熵，本文设计与实现一种高效的最相似古文句子检索算法。最后给出了自动句对齐和最相似古文句子检索的实验结果。

关键词：计算机应用；中文信息处理；古今汉语平行语料库；句子对齐；相似句子；基于实例的机器翻译
中图分类号：TP391　　文献标识码：A

Ancient Sentence Search Based on Sentence Auto-Alignment in Parallel Corpus of Ancient and Modern Chinese

GUO Rui, SONG Ji-hua, LIAO Min

(Information Science and Technology College, Beijing Normal University, Beijing 100875, China)

Abstract: Along with the Corpus Linguistics' prosperity and development, the research on Example Based Machine Translation (EBMT) has a flourishing prospect. In this area, two problems must be solved: 1) Constructing a large-scale parallel corpus with high accuracy and speed. 2) Searching the most similar sentence with the input sentence from the huge aligned examples. This paper aimed at EBMT between ancient and modern Chinese. First, a new translation model was built which takes the length of the sentence, character information and punctuation into account at the same time. Then, a new approach for aligning bilingual sentences automatically was proposed based on genetic algorithm and Dynamic Programming. Finally, a new similarity method was given based on Chinese characters' information entropy. Experimental results showed that our methods achieved good performance.

Key words: computer application; Chinese information processing; parallel corpus of ancient and modern Chinese; sentence alignment; similar sentence; EBMT

1 引言

任何民族的发展都不能没有继承。我国大量优秀的传统文化均以古代汉语作为载体。中华文化要传承、要变革，都得要有一批人读文言文，整理古籍，研究历史[1]。

然而，由于客观历史条件的限制，人们使用的现代汉语同汉语古籍中使用的古代汉语相比，存在着显著的差异。这给非专业人士阅读古代汉语、接受华夏古文明的熏陶造成了严重的障碍。目前，大量的汉语古籍已被翻译为现代文作品。然而，由于我国历史悠久，文化遗产丰富，用文言记录的典章、制度、史料以及撰写的文学作品多到不可计数，翻译这些作品本身是一项工作周期较长、工作量极大、质量要求极高的工作，仅仅依靠少数专家学者的努力难以完成。而且，专家学者建立的理论知识体系，一般是从大量古籍中古汉语的使用现象归纳得来，由于

收稿日期：2007-04-10　定稿日期：2007-12-27
基金项目：国家社科基金资助项目(05BYY022)

作者简介：郭锐(1982—)，男，硕士生，主要研究方向为古籍信息处理和机器翻译；宋继华(1963—)，男，博士，副教授，主要研究方向为语言信息处理、知识工程；廖敏(1985—)，男，硕士生，主要研究方向为古籍信息处理和机器翻译。

人的生理条件的限制, 很难穷尽所有的语言现象, 进而导致这些理性知识难免会存在偏差。

随着科学技术的进步, 特别是计算机科学的蓬勃发展, 机器翻译已成为突破语言障碍的重要技术手段。其中, 由于语料库语言学的兴起, 基于实例的机器翻译(EBMT)得到越来越多的研究。EBMT 系统以互译对照的实例库为主要的知识源, 而互译实例的收集要比翻译规则的获取容易得多。当输入待译句子 *Input* 后, EBMT 系统在互译实例库(一般是句子级对齐)中查找与 *Input* 最相似的源语句子 S, 再模拟 S 的译文 T 生成 *Input* 的译文。

如何快速准确地构建大规模古今汉语互译实例库, 以及从大量的互译实例中检索与输入句子最相似的源句子是基于实例的古今汉语机器翻译必须解决的首要问题。本文借鉴当前共时语言双语语料库中句子对齐方法和相似源语句检索策略, 结合历时语言——古今汉语的特点, 设计与实现了古今汉语自动句对齐及相似古文句子检索算法。实验结果是令人满意的。

2 自动句对齐

2.1 语料预处理

从互联网上获得的古今汉语平行语料很难满足后续程序处理的要求, 必须进行预处理。其中存在的问题大概可以分为: (1) 篇幅很长。网上语料一般都是按著作或卷整理的, 下载或者拷贝下来的文档都在万字级, 文档过长会降低后续程序处理的效率和准确度; (2) 多余标记。网上语料一般是网页格式, 含有大量的网页标记和其他非文字信息; (3) 对齐级别。后续程序自动句对齐的起点是段对齐, 网上语料并没有达到段对齐; (4) 标点符号。网上语料中古文和今文采用的标点符号是不一致的。例如古文引号和双引号分别是『』,「」。

我们编写了预处理程序, 首先去掉多余的标记信息; 接着统一古文和今文标点符号; 然后以篇章为单位, 将长文档分成 1 000 字以内的短文档, 采用程序检测和人工后对齐的方法做到古文篇章和译文篇章间的段对齐。这是后续自动句对齐的起点。

2.2 古今汉语互译模型

自动句对齐的核心问题是如何衡量一个原文句子 S 和一个译文句子 T 之间互译性的大小, 即 T 作为 S 译文的可能性有多大。共时语言的自动句对齐算法有以下三种: 基于长度; 基于词汇; 二者结合。

基于长度的方法: 是由文献[2, 3] 提出来的, 主要是根据源语言和目标语言的句长度的强相关性, 其基本思想是长句子的译文也较长, 短句子的译文也较短。该方法已经在英法和其他同语系语言的对齐中得到了较好的结果, 例如英法、英德双语对齐。

基于词汇的方法: 利用双语词典和词汇信息来对齐句子。文献[4, 5] 分别根据双语单词的分布信息和词汇翻译模型进行了英德和英法双语句子对齐。

二者结合的方法: 基于长度的方法简单直观, 早期的对齐工作仅仅依靠长度关系却获得了惊人的成果。然而该方法却存在着以下问题: (1) 基于长度的方法缺乏鲁棒性, 它对需要对齐的语料库要求较高; (2) 缺乏错误恢复能力, 当对齐出现错误时, 如在句子长度比较相似的一段对齐文本中删除其中的任何一句话都可能带来一系列的错误; (3) 在不同语系之间对齐时, 长度对应关系并非完全可信。基于词汇的方法使用了语言知识, 因此对齐的结果更加科学。然而作为对齐基础的语言知识的错误, 很可能造成对齐结果错误膨胀; 而且由于对齐过程涉及到语言知识的处理, 对齐的时间耗费会显著增加。因此近年来语料库的句对齐往往是采用二者结合互补的方法。如香港的 DeKai wu 利用 Gale 和 Church 基于长度方法实现了英语和繁体汉字之间的对齐, 还利用了日期、机构名等特殊词表将长度方法与词汇方法结合用于句子对齐[6]; Haruno and Yamazaki 利用此方法进行了英语和日语的句子对齐[7]。

这里, 用 L_s 代表古文, L_t 代表 L_s 的现代汉语译文, L_t 和 L_s 均可能包含多个句子。L_t 作为 L_s 译文的可能性大小用互译得分 *TraScore* 来表示。借鉴共时语言对齐方法及古今汉语翻译自身特点, 我们认为影响 *TraScore* 的因素有句子长度、汉字字形、标点符号。

2.2.1 句子长度因素

在句对对齐之间的无关性假设下, 利用贝叶斯公式进行推导, 可以得到只考虑长度因素下, L_s 和 L_t 互译可能性大小的计算公式[8], 我们把它记作 L_s 和 L_t 的句长互译得分 *LenTraScore*:

$$LenTraScore = Prob(M(L_s, L_t)/|l_s|, |l_t|) \quad (1)$$

其中古文 L_s、现代汉语译文 L_t 的长度分别记作 $|l_s|$, $|l_t|$, 长度的单位可以是汉字或者词, 对古

代汉语而言，单字词占据了绝大多数；另一方面，如果引入现代汉语分词机制，那么分词引入的细微错误蔓延到自动对齐阶段，可能会带来对齐准确度的大幅度降低。所以我们按字为单位来统计长度信息。古文中的句读信息是后人为了阅读传播的方便加进去的，所以在计算长度的时候，古文和译文的标点都不计算在内。例如：古文“鄒忌脩八尺有餘，身體昳麗。”的$|l_s|=11$；其译文“邹忌身高六尺有余，仪表俊美。”的$|l_t|=12$。$M(L_s, L_t)$表示互译句子间的对齐模式，$LenTranScore$表示在给定句子对L_s和L_t下，它们属于对齐模式M的概率。

上式可以通过贝叶斯公式转化为可计算模型如下：

$$Prob(M(L_s, L_t)/|l_s|, |l_t|) = \frac{Prob(d/M(L_s, L_t))Prob(M(L_s, L_t))}{Prob(d)} \quad (2)$$

其中d是古文L_s和译文L_t之间的距离度量，d的计算考虑到了句子组长度差异和全部语料库均值和方差[2]：

$$d(|l_t|, |l_s|) = (|l_t| \vdash |l_s| \times c) / \sqrt{|l_s| \times s^2} \quad (3)$$

其中$c=\sum_{A\in Corpus}|l_t| \Big\backslash \sum_{A\in Corpus}|l_s|$，表示语料库中译文总字数和古文总字数的比值，即$L_s$中的任一字符在译文$L_t$中出现的平均字符个数（期望）；$s^2$是$c$的方差。我们对人工对齐的20 296对句对进行了统计，得到c约为1.81，s^2约为0.36。

$P(M(L_s, L_t))$是句子对齐模式的先验概率，对人工对齐的20 296对句对的对齐类型进行统计得到结果如表1（从中可以看出平行语料库中对应文本长度的比值和1接近，而基于长度的对齐算法在确定“1∶1”对齐模型上性能最好[9]，也证明了这种方法在古今汉语平行语料库中实施的科学性）。

表1　基于人工对齐的20 296对句对的对齐类型统计结果

对齐类型	1∶1	1∶2	2∶1	2∶2	1∶0	0∶1
对齐概率	0.892	0.042	0.010	0.036	0.010	0.010

2.2.2　汉字字形因素

为弥补基于长度方法的不足，我们从汉字角度来衡量L_s和L_t互译的可能性。其出发点是互为译文的古今汉语中应该具有更多相似的汉字，且这些汉字出现的先后次序应大体相同。为了消除古文中大量繁体字的影响，在计算共现汉字之前先把古文中繁体字简化。同样也可以译文中简体字繁化，但我们并没有这么做，主要因为一个繁体字确切地对应一个简体字，而一个简体字可能对应多个繁体字，简体字繁化会带来歧义消解问题。例如古文L_s“城北徐公，齊國之美麗者也。”简化后变为L_s'：“城北徐公，齐国之美丽者也”。其译文为L_t：“城北徐公是齐国的美男子。”

编辑距离：两个字符串之间的编辑距离是指仅通过插入、删除、替换操作，把一个字符串变成另一个字符串所需的最小操作数目。编辑距离考察了两个字符串间的共同字符，也考察了共同字符出现的先后次序。基于词典的对齐算法也利用了动态规划算法，只不过求句对的最大概率转换成求解每一个句对的评价函数，而这里采用的评价函数就是编辑距离。

L_s和L_t的字形互译得分为：

$$CharTraScore(L_s, L_t) = 1 - EditDis(L_s', L_t) / \mathrm{Max}(|L_s'|, |L_t|) \quad (4)$$

$EditDis(L_s', L_t)$为L_s'和L_t之间的编辑距离，$|L_s'|$为L_s的长度。

2.2.3　标点符号因素

语料中的标点符号虽然在计算长度的时候被略去，但是对于句子对齐的判断具有重要的作用。通过对人工对齐的句对的考察，发现对齐句对中，原文和译文的最后一个标点绝大多数是一样的，这很大程度上可能源于古文的翻译模式，都是依据古文逐句翻译，然后在末尾加上一样的句读符号。所以句末标点就作为我们第三个计算互译可能性的参考因素，记作标点互译得分$PuncTraScore$。如果古文和译文句末标点符号一样，则$PuncTraScore=1$；否则$PuncTraScore=0$。

2.2.4　互译模型

对于任意的古文L_s，L_s的现代译文L_t，二者的互译得分是上述三个得分的加权平均：

$$TraScore = \alpha LenTraScore + \beta CharTraScore + (1-\alpha-\beta)PuncTraScore \quad (5)$$

其中权重α、β需要通过训练得到。

2.3 对齐算法

2.3.1 动态规划

基于2.2.4中的互译模型,采用动态规划算法实现古今汉语的自动句对齐。假设一个段落内的古文句子表示为 S_i, $i=1, 2, \cdots, s$;互译段落内的译文句子表示为 T_j, $j=1, 2, \cdots, t$。对齐的句子序列表示为:$A(k)=\{<Sk, Tk>, k\in[1, K]\}$,其中 K 是对齐句对的个数。记古文前 i 个句子和译文前 j 个句子组成的所有对齐中,对齐的最高得分 $Score(i, j)$,S_i 和 T_j 的互译得分为 $TraScore(S_i, T_j)$。动态规划的核心递推公式是:

$$Score(i,j)=\max\begin{cases}Score(i-1,j)+TraScore(S_i, null) & 1:0\text{型}\\ Score(i,j-1)+TraScore(null, T_j) & 0:1\text{型}\\ Score(i-1,j-1)+TraScore(S_i, T_j) & 1:1\text{型}\\ Score(i-1,j-2)+TraScore(S_{i-1}, T_{j-2}+T_{j-1}) & 1:2\text{型}\\ Score(i-2,j-1)+TraScore(S_{i-2}+S_{i-1}, T_{j-1}) & 2:1\text{型}\\ Score(i-2,j-2)+TraScore(S_{i-2}+S_{i-1}, T_{j-2}+T_{j-1}) & 2:2\text{型}\end{cases} \quad (6)$$

$S_{i-2}+S_{i-1}$ 中的+号表示字符串的连接,即把第 $i-2$ 个和第 $i-1$ 个古文句子顺序连接为一个整体参与计算互译得分。在将最高得分保存在 $Score(i, j)$ 的同时,将对齐类型保存在 $Trace(i, j)$ 中用于回溯求取最佳对齐路径。

2.3.2 权重计算

对齐算法的难点在于权重 α,β 的选择,单纯依靠经验很难给出可靠的权重。利用已经对齐的20 296句对,采用遗传算法来确定合适的权重。为了定义遗传算法的健壮度函数 $FittnessFunction$,先给出几个和对齐结果相关的定义。在给定的 α,β 值下,设人工对齐的句对集合为 A,程序自动对齐的句对集合为 B,那么自动对齐结果的准确率 $precision=|A$ 与 B 的交集$|/|B|$;自动对齐结果的召回率 $recall=|A$ 与 B 的交集$|/|A|$。$F=(2\times precision\times recall)/(precision+recall)$ 作为 $FittnessFunction$。

3 相似古文句子查找

如何迅速准确地从大量句对齐语料中找到和输入句子最相似的句子,是古今汉语平行语料库高级应用乃至古今汉语机器翻译的基础性环节。针对输入古文句子,采用逐句匹配,逐句计算相似度,然后相似度排序,最终输出结果的方法显然是不切实际的。这项应用和搜索引擎在本质上是相同的,都可以归结为在海量数据中,迅速找到前 N 个最相关的结果。为语料库建立索引及相应检索算法是目前满足这一应用需求切实可行的方法。考虑到实际需求,本文只讨论古文向现代文的翻译,暂不考虑现代文向古文的翻译。

搜索一般而言可以分为两大部分:索引和检索。索引是检索的基础,检索是索引的导向,二者是辩证统一的。检索策略很大程度上决定着索引机制,即需要对哪些信息进行索引及如何索引。我们的需求是找到和输入古文最相似的古文,相似程度的大小仍然用编辑距离来衡量。

本文的检索机制分为两步:第一步是从语料库中检索和输入古文 $Input$ 具有相同汉字的句子,得到候选结果集 $Result'$;第二步是计算 $Result'$ 中每个句子和 $Input$ 的语句相似度,得到最相似的 N 个句子作为最终结果集 $Result$。如果 $Result'$ 中的结果太多,肯定会加重第二步的计算和排序负担;如果 $Result'$ 中的结果太少,又可能遗漏掉准确的候选结果。需要平衡查全率和查找性能。

我们认为,出现在很多句子中的高频汉字,对古文句子的区分作用要远远小于只出现在少数句子中的汉字。在已对齐的20 296个句对中,“之”字在其中的5 753个句对中出现;“扈”字在其中的53个句对中出现。如果 $Input$ 中含有“之”、“扈”二字,那么 $Result'$ 中将至少含有5 753个古文句子。为了避免高频字查找带来的过多结果,又尽可能的不遗漏潜在相似句子。引入了汉字信息熵的定义:

$$H(ch)=lg(M/m) \quad (7)$$

其中 ch 表示一个汉字,M 表示对齐语料库中的句对总数,m 表示古文中出现了 ch 的句对数。

检索算法:

输入:待检索古文 $Input$,最终结果集的大小为 N,汉字信息熵最小阈值 D,$Input$ 中信息熵低于 D 的汉字将被剔除,不作为检索条件。

算法输出:最相似的 N 个古文句子及译文。

步骤:

(1) 确定 $Input$ 的字集合中哪些汉字的信息熵大于等于阈值,得到集合 CH;

(2) 对 CH 中的每一个汉字 ch_i, 去索引中查找包含该汉字的古文句子得到集合 Sen_i;

(3) 计算所有 Sen_i 的并集 Sen;

(4) 对 Sen 中的每个句子 S_i, 计算 $Input$ 和 S_i 之间的相似度;

$$Sim(Input, S_i) = 1 - EditDis(Input, S_i) / \mathrm{Max}(|Input|, |S_i|) \quad (8)$$

(5) 对(4)中结果按照相似度进行排序, 输出最大的 N 个候选结果及其译文。

4　实验结果

为了检验上文中提出的方法的正确性和可行性, 我们实现了古今汉语自动对齐与相似古文检索系统。本文的实验数据基于人工对齐的 20 296 对句对, 按内容分成两部分分别进行测试, 最后测试了一个总的结果, 且均为封闭测试。硬件环境为: CPU-3.0G, 内存 1G。如表 2 ~表 4 所示。

索引检索性能, 我们从对齐的语料中任意抽取出 10 句古文, 利用我们的检索算法得到最相似的 5 句古文句子, 花费的总时间记作检索的总时间, 其结果如表 5 所示。

检索的正确性：由于采用的是封闭测试, 语料库中和输入古文句子中完全相同的结果均作为 5 句古文中相似得分最高的句子返回。

实验结果表明, 我们的自动对齐方法取得了较好的效果, 基本可以满足大规模古今汉语平行语料库的建设要求。我们的检索机制的性能受语料库规模的影响较小, 可以满足将来语料库规模扩大后的检索要求。

表 **2**　基于人工对齐的《国语》**7 475** 对句对的参数结果

句长权重	字形权重	标点权重	准确率	召回率	F 值
0.292	0.457	0.251	0.993 8	0.993 6	0.993 7

表 **3**　基于人工对齐的《战国策》**12 821** 对句对的参数结果

句长权重	字形权重	标点权重	准确率	召回率	F 值
0.206	0.632	0.162	0.981 8	0.981 3	0.981 6

表 **4**　基于人工对齐的所有 **20 296** 对句对的参数结果

句长权重	字形权重	标点权重	准确率	召回率	F 值
0.117	0.704	0.180	0.991 1	0.991 0	0.991 1

表 **5**　检索最相似古文句子的时间效率

	4 125 句对	12 562 句对	20 296 句对
索引总时间	5 162 ms	17 629 ms	27 461 ms
平均索引时间	1.6 ms	1.4 ms	1.3 ms
检索的总时间	790 ms	993 ms	1 103 ms
检索平均时间	79 ms	99 ms	110 ms

5　结束语

古今汉语的机器翻译尚处于起步阶段, 建设大规模高质量的古今汉语平行语料库是基于实例的古今汉语机器翻译的一项极其重要的语言资源建设工作。本文提出的自动对齐算法会大大提高语料库建设的速度和质量; 索引检索机制则是语料库高级应用的基础。语言资源的建设是一项长期而复杂的工程, 我们提出的方法还有待于进一步的检验和修正。

近期的工作展望: (1) 文中的实验结果是封闭测试下得到的, 需扩展到开放测试中; (2) 随着标准语料规模的扩大, 用遗传算法得到的权重因子会更加准确; (3) 目前古文句子间的相似度基于编辑距离, 有待于进一步完善; (4) 索引机制和检索算法, 可以得到进一步的优化。

(下转第 105 页)

表 3　三种模型在平均每状态 4 高斯时的识别性能对比

	均匀分配模型	BIC 非均匀分配	MMI 非均匀分配
句子错误率(%)	19.83	16.62	14.89
数字错误率(%)	3.98	3.34	3.03

5　总结

本文提出了一种基于区分性的最大互信息量准则的声学模型拓扑结构优化方法，并将其与传统的基于模型似然度和复杂度的优化方法进行了对比。通过定义合理的启发性度量，我们尝试在一个训练充分的多高斯均匀分配模型的各个状态之间“交换”高斯核，从而使得最大互信息量准则得以优化。实验结果表明，由于基于区分性的准则的优化更为直接地将模型结构与模型的区分度和鉴别力联系了起来，因此，也就能够取得比单纯基于似然度的方法更好的识别性能。本文的工作主要在中文连续数字串这一较小的任务下进行，相关方法在应用到更大规模任务下后有可能遇到的运算量及推广性方面的问题，可以作为今后研究的方向。

参考文献：

[1] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle[A]. International Symposium on Information Theory [C]. 2nd, Tsahkadsor, Armenian SSR, Hungary, 1973. 267-281.

[2] G. Schwarz. Estimating the Dimension of a Model[J]. Annals of Statistics, 1978, 6(2): 461-464.

[3] J. Rissanen, Stochastic Complexity in Statistical Inquiry [M]. World Scientific Publishing Company, 1989.

[4] S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanevsky and P. Olsen. Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News [A]. In: Proc. ICASSP 1999 [C]. 1999. 37-40.

[5] 何珏，刘加. 汉语连续语音中 HMM 模型状态数优化方法研究[J]，中文信息学报，2006，(6)：83-88.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition[A]. In: Proc. ICASSP 1986 [C]. 1986. 49-52.

[7] B. H. Juang, W. Chou, C. H. Lee. Minimum Classification Error Rate Methods for Speech Recognition[J]. IEEE Trans. Speech and Audio Processing, 1997, 5(3): 257-265.

[8] Y. Normandin. Optimal Splitting of HMM Gaussian Mixture Components with MMIE Training[A]. In: Porc. ICASSP 1995 [C]. 1995. 449-452.

[9] D-P. R. Schluter. Investigations on Discriminative Training Criteria[D]. Ph.D. thesis, 2000.

(上接第 91 页)

参考文献：

[1] 李如龙. 文言 白话 普通话 方言[J]. 语言文字应用，2003，4：2-9.

[2] W. A. Gale, K. W. Church. A Program for Aligning Sentences in Bilingual Corpora[J]. Computational Linguistics, 1993, 19(1), 75-102.

[3] P. F. Brown, J. C. Lai, R. L. Mercer. Aligning Sentences in Parallel Corpora[A]. Proc. of the 29th Annual Meeting of the ACL-29[C]. 1991, 169-176.

[4] M. Kay, Martian. Roscheisen. Text-Translation Alignment[J]. Computational Linguistics, 1993, 19 (1)：121-142.

[5] S. F. Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information[A]. Proc. of the 31st Annual Meeting of the ACL-31[C]. 1993, 9-16.

[6] Dekai Wu, Pascale Fung. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition [A]. Morgan Kaufmann Publishers Inc. 1994.

[7] Masahiko Haruno, Takefumi Yamazaki, High-performance bilingual text alignment using statistical and dictionary information [A]. Association for Computational Linguisticsc[C]. 1996. 131-138.

[8] 张艳，柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报，2005，19(5)：31-36.

[9] Christopher D. Manning , Hinrich Sch tze, 苑春法，等译，统计自然语言处理[M]. 北京：电子工业出版社，2007, 292-309.