第31卷 第16期 计 算 机 工 程 2005年8月
Vol.31 № 16 Computer Engineering August 2005
·基金项目论文· 文章编号：1000—3428(2005)16—0045—03 文献标识码：A 中图分类号：TP391

基于双语词典的汉英词语对齐算法研究

邓 丹，刘 群，俞鸿魁
（中国科学院计算技术研究所数字化室，中国科学院研究生院，北京 100080）

摘 要：研究利用多部人读双语词典扩充双语词典的规模来改善词语对齐质量。介绍了一个在Ker算法基础上用双语词典进行汉英词语对齐的算法。提出了对齐窗口的概念，通过在对齐过程中设置对齐窗口，可以找到多对多的词语对应。
关键词：词语对齐；对齐窗口；人读双语词典；机读双语词典

Research of Chinese-English Word Alignment Algorithm Based on Bilingual Dictionary

DENG Dan, LIU Qun, YU Hongkui
(Digital Technology Lab, Institute of Computing Technology, Chinese Academy of Sciences,
Graduate School of Chinese Academy of Sciences, Beijing 100080)

【Abstract】A large scale of bilingual dictionary enlarged by integrating several human-readable bilingual dictionaries is the main course to improve the result. A Chinese-English word-alignment algorithm based on bilingual dictionary is introduced. It is inspired by Ker's method. The concept of "alignment window" is proposed. By setting alignment window, many-to-many word alignments can be found.
【Key words】 Word alignment; Alignment window; Human-readable bilingual dictionary; Machine-readable bilingual dictionary

通过多部人读双语词典可以扩充机读双语词典的规模。人们希望使用大规模双语词典能够克服一般用双语词典进行词语对齐覆盖率不高的问题。在Ker算法的基础上我们构造了一个新的词语对齐系统，舍弃了同义词词典，保留了双语词典。利用双语词典、位置、词性信息进行词语对齐。我们改进了Ker对相对位置偏移的计算方法，并在贪婪算法之前，通过词语相似度选择对齐锚点来改善贪婪算法的对齐质量。在贪婪算法过程中，提出了“对齐窗口”的概念，找到了多对多的单词对应。

1 用双语词典进行词语对齐

1.1 词语对齐的定义

互译句子对中，源句子的一个单词和译文句子中的一个单词形成了一个连接。词语对齐就是要在两个句子的单词形成的所有连接的集合中寻找一个子集，使得子集中的每一个连接都是在句子中有互译关系的连接。

1.2 多部人读双语词典扩充机读双语词典

扩充双语词典规模以改善词语对齐质量的想法是自然而然的。但人工开发大的双语词典费时费力（这就是为什么词典规模有限）；从双语语料库中抽取词典需要好的学习算法，这跟词语对齐一样难，且易产生垃圾信息，同时要受语料库的影响。

到目前为止，已经有很多双语电子词典。但这些词典往往格式不规范，被称为人读双语词典。词语对齐用的双语词典实际上是格式规范的机读双语词典。刘群[5]等人将6部人读汉英和英汉双语词典转化为了机读双语词典，计中英文词形和词性（不计大小写），扩充前核心词典共有词条48 361条；扩充后的词典有460 207条。

由于这些人读词典都是面向通用领域的，因此最终得到的也是一部面向通用领域的机读双语词典，这对进行词语对齐非常有利。

1.3 对齐前的预处理

首先进行词性标注，中文用ICTCLAS2.0，标注集是ICTPOS；英文用Eric Brown标注器，采用WSJ标注集。通过禁用词词性表将难以找到对齐的虚词滤出，以免干扰其他单词的对齐。用Wordnet附带的词形变换表对英文词进行词形还原，查不到的单词用规则还原，还原歧异通过查双语词典的英文词表定夺。另外，中文句中的人名用拼音代替，数词用阿拉伯数字代替。

1.4 词语相似度的计算

Dice系数可以计算两个字符串的相似度：

$$Dice(s_1,s_2)=\frac{2*comm(s_1,s_2)}{leng(s_1)+leng(s_2)} \quad (1)$$

其中，comm (s_1,s_2)是 s_1,s_2 中相同字符的个数，leng(s_1), leng(s_2)是字符串 s_1,s_2 的长度。

对中文，一个汉字算2个字符；对英文，comm(s_1, s_2)用最大公共子串（连续的）长度计算。

利用双语词典通常用Dice系数计算连接(s, t)的词语相似度值sim(s, t)。它是连接(s, t)中一个词在词典中所有义项与另一个词的Dice系数的最大值。双语词典可以双向利用，用中文和英文的都可计算Dice系数；将s, t看作英文字符串可计算Dice(s, t)，最终取这3种Dice系数的最大值作为sim(s, t)。

基金项目：国家重点基础研究发展规划项目（G1998030504-02）
作者简介：邓 丹(1979—)，女，硕士，主研方向：自然语言处理；刘 群，副研究员；俞鸿魁，硕士生
定稿日期：2004-06-14 **E-mail：**ddan@ict.ac.cn

1.5 相对位置偏移的计算

s, t 是源文中和译文中的第 i 个和第 j 个单词，(i_L, j_L), (i_R, j_R)是 s 两边离 s 最近的已经对齐的连接的位置，那么 Ker 的相对位置偏移 rd 为：

$$rd(s,t) = \min(|d_L|, |d_R|) \quad (2)$$

其中 $d_L=(j-j_L)-(i-i_L)$, $d_R=(i-i_R)-(j-j_R)$

rd 值越小单词对齐的可能性越大。由于 rd 是相对的，所以 rd 可能是以相距很远的词对计算的，因此在计算 rd 时加了一个阈值 THR（取 5），考虑了绝对位置偏移量：

$$d_L=\begin{cases}(j-j_L)-(i-i_L), if\ (|i-i_L|<THR)\\ i-i_L, else\end{cases} \quad (3)$$

对 d_R 的计算也作同样的处理。

Ker 通过在句子的两端加空标志，将空标志形成的连接作为初始对齐连接计算 rd 值。除了将句子两端的空连接作为初始对齐连接以外，还通过词语相似度值 sim，将那些 sim 大于阈值 I（I 取 0.5），且其它与 s, t 相关的连接的 sim 值都不大于 I 的连接作为初始对齐的连接，使得一开始计算 rd 时，句子中有真正对齐连接作为锚点。

1.6 用 Bootstrap 方法训练词性转移概率

所谓词性转移概率，就是一个词性翻译成译文中某种词性的概率。原文词性翻译成译文中不同词性的概率不同。在标有词性的词语对齐的语料库中统计条件概率，就得到词性转移概率表。

我们用 Bootstrap 的方法在句子对齐的语料库上训练出词性转移概率表。方法是初始用不使用词性信息的词语对齐方法对句子对齐的训练语料库进行词语对齐，得到准确率很高的对齐结果。然后在对齐结果中统计词性转移概率。将统计出的词性转移概率加入到词语对齐系统中，改善对齐结果，使得在准确率较高的情况下，召回率上升（准确率、召回率在词语对齐的测试语料上测得），又用新的词性转移概率表改善对齐结果，如此迭代，直到词语对齐的结果不再改进为止，这时就得到了最终的词性转移概率表。

1.7 计算词语对齐概率的方法

用独立性假设，词语对齐的概率为：

$$\begin{aligned}\Pr ob(s,t) &= \Pr ob(s,t \mid sim(s,t))\\ &\times \Pr ob(s,t|rd(s,t))\\ &\times \Pr ob(s,t \mid pos_s, pos_t)\end{aligned} \quad (4)$$

其中 $\Pr ob(s,t|sim(s,t))=sim^3(s,t)$ (5)

sim(s, t)的值在 0~1 之间。对它取 3 次方幂作为词语相似度的概率，3 实际上是经验值。试验表明这样能较好抑制用 Dice 系数计算词语相似度的噪声，得到较好的准确率和召回率。对于 $\Pr ob(s,t|rd(s,t))$，rd 用改进的方法计算。概率值用 Ker 的概率表(表 1)，增加了 rd>3。 $\Pr ob(s,t|pos_s,pos_t)$ 是在训练中直接得到的。

表 1 相对位移概率表

条件	rd=0	rd=1	rd=2	rd=3	rd>3
概率值	0.26	0.11	0.07	0.04	0.01

1.8 利用对齐窗口得到多对多的词语对齐

预处理后，对两个句子的单词做笛卡儿集，得到带有词语、词性、位置信息候选连接集。

Ker 采用贪婪算法挑选对齐词对，过程是：

计算出候选集中每一个连接的词语对齐概率后，从中选出概率值最大连接(s,t)，加入到对齐的连接集中去，同时从候选连接集中删除出现了 s 或 t 的所有连接。更新候选连接集中连接的 rd 值和词语对齐概率值，再次从中挑选，直到选出的连接的词语对齐概率值不大于阈值或候选连接集为空。

Ker 的算法中每次选出一个对齐概率值最大的连接(s,t)加入到对齐连接集后，从候选集中删除所有出现 s 或 t 的连接，那么最终选出来的对齐连接只能都是一对一的连接。若要得到多对多的连接，就不能在选出一个最好的连接(s,t)后不加分辨地将所出现 s 或 t 的连接删去。

通过设置“对齐窗口”的方法，在每次选出最好的连接(s,t)后，有条件地直接选出或者暂且保留跟 s,t 有关的部分连接。

“对齐窗口”的定义是：已知(s,t)是贪婪算法中选出的对齐概率值最大连接，如果 s,t 在句子中存在多对多的连接，那么这些连接中的单词必在 s，t 附近的 2n 个单词形成的窗口中，不可能在窗口之外，n 是一个常整数。对齐窗口基于多对多的单词对应中的单词在句中往往有隔得很近的现象。

在图 1 的例句(e, f)中，要开 n＝2 的对齐窗口，假设刚刚选出了 (法制, legal)(图中粗体表示)，加入对齐连接集，并从候选集中删除，此时，对齐窗口就是图中的阴影部分。称 e 中的为“法制”窗口，f 中的为“legal”窗口。对于候选集中存在连接来说，“法制”跟“legal”窗口中的单词形成的连接可能是对齐的连接(例如(法制，system))，要通过考察才决定去留；而跟“legal”窗口之外单词形成的连接（例如(法制，establishment)）被认为不可能是对齐的连接，直接从候选集中删去。“法制”窗口的作用也一样。

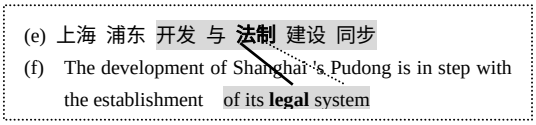

图 1 对齐窗口

当 n!=0 时，就可以找到最多 1 对 2n+1 的单词对应。找到了一对多，就自然可以找多对多的对应。当 n＝0 时，就只能找一对一的对应，这时对齐窗口就不起作用了。

整个对齐算法过程如下：

(1) 读入一对中英文句子。

(2) 进行切分标注，过滤禁用词，词形还原等预处理。

(3) 对没有被过滤掉的中英文单词做笛卡儿集作为候选连接集。计算候选连接集中每一个连接的 sim 值，并从训练好的词性条件概率表中查出连接的词性条件概率值。

(4) 设置对齐连接集为句子两端形成的两个空连接的集合。选择 sim 值大于阈值 I(取 0.5)，且其连接中的单词与其它单词形成的连接的 sim 值均不大于 I 的连接作为锚点加入到对齐连接集中，删除候选连接集中所有跟锚点单词有关的连接。

(5) 贪婪算法选择对齐连接：计算候选集中所有连接 rd 值（用改进的方法计算），结合(3)中已有的 sim 值和词性条件概率，根据公式(4)计算连接的对齐概率，从中选择对齐概率最高的连接(s, t)。若(s, t)的单词对齐概率不高于阈值 T（如 9e-5）或者候选集合为空进入(6)；否则将(s, t)插入到对齐的连接集合中，删除候选集中的(s, t)，然后分情况作如下处理：1) 若 sim(s, t)高于阈值(如 0.9)，则认为 s, t 不可能与其它单词形成对齐连接，删除候选集中所有跟 s, t 相关的连接，直接回到(5)，继续选择。2) 否则对 s，t 各开 n＝N 的对齐窗口(如 N=2)。删除 s 与 t 窗口之外，t 与 s 窗口之外的单词形成所有连接。考察 s 与 t 窗口中的单词 t’形成的连接(s, t’)是否是对齐的连接：若 sim(s, t’)大于等于阈值 A（如 0.7），而候选集中其他 (s‘, t‘)的 sim (s‘, t‘)均小于阈值 B（如 0.4），那么(s, t‘)是对齐连接，加入到对齐连接集，删除所有与 t‘有关的连接；否则若 sim(s, t’)

小于阈值 A，且候选集中有其它 (s ‘ , t ‘)使 sim (s ‘ , t ‘)大于等于阈值 A，则删除 (s, t’)。用同样的阈值 A 和阈值 B 考察 t 与 s 窗口中的单词形成的连接的去留。

(6) 算法结束，输出对齐的连接集合。

2 试验和结果分析

我们在规模为 10 375 对的新闻语料上用 Bootstrap 的方法训练得到词性转移概率。训练时计算准确率和召回率所需要的词语对齐的语料是从这 10 375 对的新闻语料中抽取 200 对手工做词语对齐得到的。同时在这 200 对句子上还调节算法的各阈值及经验值（如计算 sim 概率用到 sim 的 3 次幂，窗口大小取 2），使这 200 句的对齐结果最好。这些都调好以后，系统就可以对其它语料进行词语对齐了。

系统训练好后，在另外一个语料库中抽取 650 对句子做手工词语对齐作为测试集，其中共有对齐连接 16 022 个，中文句子平均单词数是 24.8，英文句子的平均单词数为 34.5。

在这个 650 对句子上得到了以下试验结果(rd 是相对位置偏移，POS 表示词性因素)，见表 2。

表 2 词语对齐试验结果

算法采用的因子	初始	准确率	召回率	F 值
1 基于核心词典	锚点	0.815	0.371	0.510
2 核心词典＋rd(改进)	锚点	0.855	0.387	0.533
3 核心词典＋rd(改进)＋POS	锚点	0.860	0.455	0.595
4 基于扩充词典	锚点	0.830	0.440	0.575
5 扩充词典＋rd(Ker)＋POS	锚点	0.840	0.600	0.700
6 扩充词典＋rd(改进)＋POS	空	0.827	0.620	0.709
7 扩充词典＋rd(改进)＋POS	锚点	0.840	0.629	0.720

试验 1 与 4 只是使用的词典规模不同，4 中加入人读词典的信息使词典规模扩大到后，召回率大幅度上升 6.9%，准确率没有下降。因为大词典对语料的覆盖面也大。在其它条件相同的情况下，从试验 7 和 3 的比较也看到：仅仅扩大双语词典的规模在带来召回率大幅度上升的同时，准确率方面有所下降，分析其原因是，随着词条数的增加会出现更多的对齐冲突（歧义），可能会降低对齐的正确率。

加入相对位置偏移概率和词性转移概率以后，词语对齐的准确率和召回率均上升（2，3 与 1 比较），因为位置和词性信息不仅进一步提供了词语对齐的证据，而且能帮助抑制用双语词典 Dice 系数计算词语相似度所带来的噪声。

有趣的是试验 4 和 3 的比较：试验 3 的召回率和准确率均比试验 4 好。这说明仅仅增加词典规模对对齐结果的改善还不如在对齐时增加位置因素和词性因素对结果的改善大。但是试验 3 也表明双语词典的规模不大时，对齐的召回率确实不高，只有 45.5%。

试验 5 和 7 的比较以及 6 和 7 的比较可以看出：对 Ker 的计算 rd 方法的改进及加入锚点，均使对齐的准确率和召回率有提高。

最终用我们的单词对齐系统在测试集取得准确率 84.0% 的条件下 62.9%的召回率（表 2 中的试验 7）。因为测试集中的句子都比较长，对齐难度比较大，所以这个结果还是比较令人满意的。最后，为考查对齐窗口在对齐中的作用，使用表 2 中的试验 7 相同的设置，仅调节对齐窗口中的 n 值的大小，得到表 3 所示的试验数据。

当 n＝0 时，对齐窗口不起作用，只能找出一对一的词语对应。当使用对齐窗口以后（n 取大于 0 的整数值），准确率略有提高，同时召回率能提高 4~7 个百分点（n=1, 2, 3），这说明开对齐窗口能在保证准确率不下降的情况下，找到更多的对齐。对齐窗口使得以前只考虑一对一的连接时，直接被删掉的一对多或者多对多的连接被找回来了一部分。还可以看到，n 值并非越大越好，n 越大，准确率会持续下降，同时召回率的上升也变慢。n 越大通过对齐窗口中保留下来的连接越多，造成的歧义也会增多，所以准确率下降。n 取 1，2 或者 3 是比较好的，在算法中取的是 2。词语对齐结果如图 2 所示。

表 3 对齐窗口在词语对齐中的作用

窗口	准确率	召回率	F 值
n=0	0.835	0.574	0.680
n=1	0.841	0.614	0.710
n=2	0.840	0.629	0.720
n=3	0.838	0.635	0.722
n=7	0.828	0.646	0.726
n=15	0.815	0.654	0.725
n=20	0.808	0.654	0.723

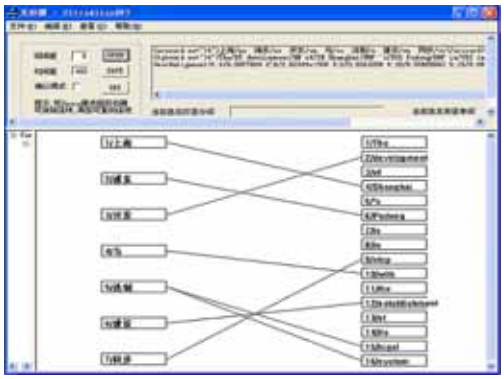

图 2 词语对齐结果

3 结束语

本文主要研究了扩大双语词典规模对对齐结果的改善情况，介绍了利用双语词典进行词语对齐的方法。它在 Ker 的算法基础上，改进了 Ker 的相对位置偏移的计算方法，加入了对齐锚点。并改进了 Ker 的贪婪算法，通过在贪婪算法中设置对齐窗口，找到多对多的单词对应。我们的算法利用大规模双语词典结合词性条件概率和相对位置偏移使得在测试集下，取得在准确率 84.0%的条件下 62.9%的召回率。

虽然随着词条数的增加会出现更多的对齐冲突（歧义），可能会降低对齐的正确率，但实验证明，用多部人读双语词典可以扩充词语对齐所用的双语词典（机读双语词典）的规模，使对齐召回率有较大幅度的上升。如果在扩大规模的同时结合多个对齐因子，可以更好地提高对齐质量。

参考文献

1 Brown P F, Della Pietra S A, Della Pietra V J, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics, 1993, 19(2): 263

2 Melamed I D.A Word-to-word Model of Translation Equivalence[C]. In: Proc. of the 35th Conference of the Association for Computational Linguistics, 1997: 490-497

3 Ker S J, Chang J S. Align More Words with High Precision for Small Bilingual Corpora[J]. Computational Linguistics and Chinese Language Processing, 1997, 2(2): 63-96

4 Huang Jinxia, Key-Sun Choi. Chinese-Korean Word Alignment Based on Linguistic Comparison[C]. In: Annual Meeting of the Association for Computational Linguistics, 2000: 392-399

5 刘 群，张 彤. 汉英机器翻译系统扩充词典的建造[C].机器翻译研究进展（全国机器翻译研讨会论文集）,北京：电子工业出版社，2002: 25-33