第 16 卷 第 1 期
1997 年 2 月

情 报 学 报

Vol. 16, №1
February, 1997

信息处理技术

基于词典和统计的语料库词汇级对齐算法[1)]

刘小虎 吴 葳 李 生 赵铁军 蔡萌
(哈尔滨工业大学计算机科学与工程系,哈尔滨 150001)
鞠 英 杰
(黑龙江大学信息管理系,哈尔滨 150080)

摘要 语料库词汇一级的对齐,对于充分发挥语料库的作用意义重大。本文对汉英句子一级对齐的语料库,提出了借助于词典和语料库统计信息的有效的对齐算法。首先利用词典的词的译文及其同义词在目标语中寻找对齐;其次利用汉语词汇与英语单词的共现统计信息以最大的互信息寻找对齐词汇以及相邻短语。实践证明该方法是行之有效的。

关键词 语料库 词汇级对齐 共现概率

Aligning Algorithm for a Corpus at Word Level Based on Dictionary and Statistics

Liu Xiaohu, Wu Wei, Li Sheng, Zhao Tiejun and Cai Meng
(*Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin* 150001)
Ju Yingjie
(*Department of Information Management, Heilongjiang University, Harbin* 150080)

Abstract Aligning the bilingual corpus at word level is very important to take the advantages of corpus. This paper presents an efficient aligning algorithm for a corpus aligned at sentence level, using the lexical information and statistic information. First, the information of dictionary and thesaurus is used. Second, the mutual information between Chinese words (or adjacent phrases) and English words (of adjacent phrases) is used. Our experiments has proved this method to be effective.

Keyword corpus, align at word level, probability of concurrence

1 引 言

近年来,语料库方法越来越受到人们的重视,同时人们也不断地认识到不经过加工的"生

收稿日期:1996 年 6 月 3 日

作者简介:**刘小虎**,男,1970 年生。哈尔滨工业大学计算机系博士生,主要研究方向为机器翻译。研制开发了达雅翻译工作站。

1) 本项研究得到国家 863 基金(863－306－03－06－3)的资助。

语料"可用性较差,能够利用的知识不多。对语料库进行多级加工,在此基础上获取各类知识的作法已被许多同行采用。语料库加工可分为:对齐,词性标注,语法标注,语义标注。其中对齐又可分为篇章级、句子级、词汇级。基于篇章、句子级的对齐算法国内外研究较多,而词汇级对齐的算法国内研究还较少。词汇级对齐的语料库对于基于语料库的翻译方法的研究,对于歧义的消解,以及对于获取词典知识、获取翻译模式意义十分重大。

人们开展了许多关于双语文本对齐的研究。例如 Chen[2],Church[3,4],Brown[5],Gale[6]。但是,大部分工作都是针对欧洲的语言对,特别是 English 和 French 之间。Church 用到的 Char-align 算法[4]比较典型,该方法寻找源语言与目标语言字符系列近似的单词作为对齐的词对。该方法主要针对欧洲语言对,因为单词往往有类似字符系列。如 government 和 gouvernement。然而,不属同一语系的汉英之间,符号集不同,所以这种 char-align 算法不适用。Shin[1]利用朝鲜语的特点,在韩语和英语之间做了词和短语级的对齐。赵铁军、李生[7]等对汉语、英语语料库的句子级对齐也作了有益的探索。

句子级的对齐算法中,基于词的个数或字符串的个数的做法,在词汇级对齐中不可能仿效,汉语词汇的长度与英文单词长度之间不存在某种映射。之所以词汇级对齐难以实现,主要存在以下几个方面的困难:

(1) 非一一对齐。汉语词汇与英文单词之间并非一一对齐,存在着一对一、一对多、多对一、多对多、对空、空对等如此复杂的对齐方式(对空,即源语言的词没有译文;空对,即译文中的内容没有源文对应)。

(2) 词的歧义。源语言词汇在翻译时,有多种译文,还可能有词性转换。

(3) 分离的短语。发现并确定分离的短语的困难,使对齐分离短语更为不易。

(4) 不同的词序。源语言与目标语言之间的词序不一致,使得利用词的先后顺序进行对齐不大可能。

(5) 灵活的译文。人工翻译的句子往往在理解的基础上意译,对词汇的翻译绝不局限于词典的解释,这使得按词典译文匹配来对齐效果不明显。

可以说,无论是上述的哪种情况,之所以带来困难,归根结底是缺乏对句子的分析。然而,非常好的分析器不易得到,而且对齐后的语料库也主要是对分析、生成提供支持。为不陷入这种互相依赖的循环,在不经过分析的语料库中进行词汇对齐是很有意义的。本文针对已经进行过句子级对齐的汉英双语语料库做词汇级的对齐研究。

本文对齐算法的主要思想是:首先利用词典的词的译文及其同义词在目标语中寻找对齐,即基于词典对齐。其次,如果利用词典对齐失败,利用从语料中统计得到的汉语词汇、相邻短语与英语单词、相邻短语的共现统计信息,用最大的互信息寻找对齐词汇,即基于统计的对齐。

作为第一步,我们只针对实词进行对齐研究,具体说只针对名词、及物动词、不及物动词、形容词、量词、数词、指代词。

2　词典、统计模型

2.1　英汉词典、类义词典、语料库

所用到的知识包括:英汉词典,汉语类义词典,汉英语料库。

所使用的英汉词典有 8 万词条,以英词单词为人口,得到该单词的汉语译文。所有译文作

为候选的对齐词汇。

汉语类义词典(thesaurus)是由清华大学黄昌宁教授提供的。根据汉语词汇的语义,将汉语词汇进行四级分类。由于一些词可能存在多个义项,每个义项的同义词也会不同。当同义词用于对齐时,词的义项是不可知的,故将该词所有义项下的同义词都作为候选的对齐词汇。类义词典由如下的记录组成:汉语词条,义项表示符,指向下一个同一义项的记录。类义词典在结构上有三个特点:(1)同一义项的记录组成一个循环链表,从任何一个记录开始,能够找到该义项下所有的同义词;(2)词条按字典序自小到大排序,便于二分查找;(3)若单词有多个义项,则组成的多个记录相邻存放。举例如下:

好看 Eb^{300101}→华美 $Eb^{3001010}$→美观 $Eb^{3001010}$→美丽 $Eb^{3001010}$→顺眼 $Eb^{3001010}$→好看 Eb^{300101}。这里,义项标记为 Eb^{300101} 的五个记录组成了一个循环链表。

所使用的汉英语料库为汉语、英语间句子级对齐的例句库,没有词性、句法标注,只有句子之间的分隔标志。该语料库有 15 万句对,我们将从中获取统计信息,然后又用于对其进行词汇对齐标注。

2.2 统计模型

我们作如下假设:

假设 1:英语单词 E 若能翻译成汉语词汇 $C1, C2, \cdots, Cm$,则 E 与 $Ci(i=1, \cdots, m)$ 之间的互信息 MI 较大,且若 $MI(E, Ci)$ 最大,则 E 最有可能翻译为 Ci。

假设 2:$C1, C2, \cdots, Cm$ 不会在一个句子同时出现一次以上。

假设 3:$C1, C2, \cdots, Cm$ 的多个义项,不会在汉语句子中同时出现一次以上。

假设 1 是做统计的基础,我们认为若 E 译为 C,则 E 与 C 的互信息最大。假设 2 和假设 3 主要目的是:在用词典对齐时,汉语句子中没有一个以上的 E 的解释或同义词,即假设不存在对齐歧义。

这些假设并不是总能满足,这也是导致错误的一些原因,但是在作了这些假设之后,才能给出能有效指导统计的模型。换句话说,作这些假设是有必要的,有利的。实际上,E 在同一个句子中有两种以上的翻译的可能性也极小。

从语料库中,我们要统计单词出现的次数、英文单词与汉语词汇共现的次数,由此计算英文单词与汉语词汇之间的互信息:

$$MI(C,E) = \log \frac{P(C,E)}{P(E) * P(E)} \approx \log \frac{M * f(C,E)}{f(C) * f(E)}$$

其中,P 表示概率,f(C)表示句子中出现 C 的句子数,f(E)表示句子中出现 E 的句子数,M 表示语料库中句子的总数,是一个常量。$P(C) = f(C)/M$,$P(E) = f(E)/M$。

对于给定的单词 E,与汉语句子中的 m 个词 $C1, C2, \cdots, Cm$ 的互信息为 $MI(C1, E)$,$MI(C2, E)$,…,$MI(Cm, E)$。为表达方便,不妨假设 $MI(C1, E) >= MI(C2, E) >= \cdots >= MI(Cm, E)$。

由于 C1 与 E 的互信息最大,我们自然选择 C1 作为 E 的译文。但是,由于存在数据稀疏问题,只用互信息判断还是不够的。例如:词 C 和单词 E 在整个语料库中各出现一次,并且在同一个句对中。设整个语料库中有 M 个句子,那么 $MI(C, E) = \log M$,显然超过了互信息的阈值 T。但此时选择 C 作为 E 的对齐词是不可信的,极有可能导致错误。为此还必须引入 T 打分机

制(t_score)来弥补互信息的不足。t_score 的定义如下:

$$t\_score=\frac{P(C,E)-P(C)*P(E)}{\frac{1}{M}P(C,E)}\approx\frac{M*f(C,E)-f(C)*f(E)}{M\quad M*f(C,E)}$$

当 t_score > θ时,选择 C 作为 E 的译文。如对上述数据稀疏问题,t_score = 1 - 1/M < θ(θ 的经验值为 2.0),所以不选择 C 作为 E 的对齐词汇。

因此,对齐的判断标准是:若 C 为使 MI(C,E)最大的汉语词汇,只有当 t_score > θ时,才选择 C 作为 E 的对齐。

3　对齐算法

设英语句子 ES = E1, E2, …, Em,汉语句子 CS = C1, C2, …, Cn, Ei 为英语单词或相邻短语,Ci 为汉语词汇。

我们从英语到汉语对齐。由于 Ei 可能是英语短语,所以该算法能够做到英语相邻短语的对齐。

统计算法比较简单,只需统计每个词汇、单词以及汉语词汇和英语单词共现的次数。假设经过统计,MI(C,E)都已经计算出来了,对每个句子的对齐算法的过程描述如下:

```
1  切分汉语句子 CS = C1, C2, …, Cm,合并英词句子的短语,结果为 ES = E1, E2, …, En
   for(i=0;i<n;i++) {
2    当 Ei 为实词或相邻短语时,查英汉词典,其解释为 {I1, I2, …, Il },共 l 项
     for(j=0;j<l;j++) {
3      if(Ij 在 CS 中出现) {
4        标记与 Ei 对齐的汉语词汇
         break;
         }
       else {
5        查 Ij 的同义词 {T1, T2, …, Tk },共 k 项
6        若某一 Tk 在 CS 中出现,标记与 Ei 对齐的汉语词汇
         break;
         }
     }
   if(Ei 没被对齐) {
7      计算最大的 MI(C,E)
8      计算 t_score
9      if(t_score > θ)
10       标记与 Ei 对齐的汉语词汇 Cf, Cn
       else
11       Ei 对齐失败
       }
   }
```

算法中的第 1 步是对英语句子预处理，包括形态还原、合并相邻的短语。第 3、6 步判断字符串 W 是否在 CS 中出现。首先试图与 CS 的切分结果一致，即 W 应是 Ci 或多个 Ci 的组合。若不能满足这一要求，就直接在 CS 中匹配 W，而不考虑原来的切分结果。若在 CS 中匹配成功，相应地修改原来的切分结果。例如，ES：it is only shallow people who believe what John said；CS：只有 \浅薄 \的 \人才 \会 \相信 \约翰 \说 \的话 \。存在切分错误。在对齐"people"时，CS 切分结果中没有词是"人"、"人们"或"人民"，不考虑切分。CS 中包含"人"，修改原来的切分为"只有 \浅薄 \的 \人 \才会 \相信 \约翰 \说 \的 \话。"，并标记"人"与"people"对齐。

算法中的第 3 步，由于 Ij 可以是分离的形式，如 salute 的解释"向…致敬"，所以算法也能对齐部分汉语分离短语。

下面举几个例子说明该对齐算法。

例 1：直接利用词典对齐

句子"Reach down that book，please.→请您把那本书拿来来。"

经过第 1 步，ES＝Reach down \that \book \，\please \

CS＝请 \你 \把 \那 \本 \书 \拿下来。

第 2 步查词典，结果为

Reach down \\拿下来，that \\那，book \\书，please \\请

满足条件 3，经第 4 步，"Reach down，book，please"分别对齐到"拿下来，书，请"。

例 2：利用类义词典对齐

句子"They held his behaviour up as a model for the others.→他们推举他的行为作为其他人学习的楷模。"

第 2 步查词典，得到"model"的解释为"模型，榜样，样板…"，未包含在汉语句子中。第 5 步查出"榜样"的同义词有"榜样，楷模，表率，师表"。经第 5 步，"model"正确地与"楷模"对齐。

例 3：利用统计信息对齐

句子"He had time to study Chinse after entered the university.→上大学后他有时间读中文。"

第 2 步查词典，得到"study"的解释为"〈1〉学习，研究，调查，分析；〈2〉考虑，努力，细想[看，察]；〈3〉学科，研究科[项]目，论文"。没有一项解释包含在汉语句子中。第 5 步查出"学习"的同义词有"学习，攻，就学，求学，上学，深造，习，向学，修，修业，学"。其他解释的同义词也不包含在 CS 中，因此，用词典对齐失败。MI(读，study)最大(f(读)＝821，f(study)＝1272，f(读，study)＝97)，而且 t_score＝3.3＞θ所以选择"读"作为"study"的对齐词汇。

4 实验及分析

为得到统计数据，我们对 6 万句对的汉英双语语料库进行了词频统计、汉语词与英语单词的共现统计，计算出了任何汉语词汇 C 与英汉单词 E 之间的联合概率 P(C，E)。P(C，E)可以理解为 C 和 E 同时出现在一个句对中的概率。如 2.2 节所述，若 C 是使 MI(C，E)最大的汉语词汇，判断 C 是否为 E 的译文标准是 t_score 是否大于阈值 θ。

我们对 100 个句子中的 1164 个英语实词或相邻短语进行了对齐算法的测试实验。实验中记录了导致错误或正确的原因。试验结果如表 1 所示。

表 1　算法适用率、正确率

对齐失败	对齐错误		对齐正确		
	由词典引起	统计方法引起	词典对齐	类义词典	统计对齐
$n1=185$	$n2=21$	$n3=33$	$n4=341$	$n5=213$	$n6=371$

可见，算法的适用率为$(n-n1)/n=84\%$。也就是说算法能对 84%的词进行对齐判断，而对 16%的词无法确定与其对齐的词或短词。有 979 个词作出正确的对齐，正确率达$(n-n1-n2-n3)/n=79.5\%$。有 $n4/n=29.3\%$的词是直接利用词典正确进行了对齐，有 $n5/n=18.3\%$的词利用同义词正确对齐。利用词典正确对齐的词占$(n4+n5)/n=47.6\%$，并不很高。而利用概率方法，使正确率提高了 $n6/n=31.9\%$。可以看出统计方法发挥了很重要的作用。

对齐失败的原因主要有两点。(1) 译文与词典的解释差别太大，而且译文与源文之间的共现概率有时很低，这时无法作出判断。这也说明了词典的信息不十分完善。有些译文在对齐中很少用上，而真实文本中出现的译文词典中却没有收入[6]；(2) 由于某些词意义很多，用法灵活，此时单词 E 译为 C1，C2，…的概率相差无几，最大的 MI(C1，E) 和次大的 MI(C2，E) 差别也不明显，对(C2，E) 计算 t_ score，也可能超过阈值。对于第二种错误原因，有两种方案有待进一步试探：(1) 考虑 C1、C2 与 E 的词性是否一致；(2) 采用约束传播算法的思想，若 E'与 C'已经对齐，且 E 与 E'之间相隔 k(k 可正可负，表示 E 在 E'前后) 个词，则 E 的对齐词汇 C 与 C'的距离大约为 k * m/n。m、n 分别为汉英句子词的个数。

为提高利用词典对齐的正确率和适用率，进一步可以使用汉英词典，从汉语出发寻找对齐的英语单词、短语。

若增加阈值 θ适用率必然降低，但正确率能提高；若减小阈值 θ适用率虽然能提高，但必然引起正确率下降。我们进一步实验了 θ变化时 n1、n3、n6 的变化情况。如表 2 所示，随着 θ的增大，错误率 $n3/(n3+n6)$ 在不断减小，而适用率却不断下降。

表 2　n1、n3、n6 随 θ变化的情况

θ	n1	n3	n6
2.0	185	33	371
2.5	214	27	346
3.0	312	19	256
4.0	405	4	180

5　结　　论

近年来，语料库方法越来越受到人们的重视。对语料库进行多级加工，在此基础上获取各类知识的作法已被许多同行采用。句子级的对齐算法国内外研究较多，而词汇级对齐的算法国内研究还较少。词汇级对齐的语料库对于基于语料库的翻译方法的研究，对于歧义的消解，以及对于获取词典知识、获取翻译模式意义十分重大。

近年来，人们开展了许多关于双语文本对齐的研究。但对非同一语系的汉语、英语之间，由于汉语词汇长度与英文单词长度间不存在某种映射，字符集也完全不同，可借鉴的方法不多。

本文对汉英句子一级对齐的语料库，提出了借助于词典和语料库统计信息的有效的对齐

算法。对齐算法的主要思想是:首先利用词典的词的译文及其同义词在目标语中寻找对齐,即基于词典对齐;其次,如果利用词典对齐失败,利用从语料中统计得到的汉语词汇、相邻短语与英语单词、相邻短语的共现统计信息,用最大或然估计寻找对齐词汇,即基于统计的对齐。

参 考 文 献

[1] Jung H·Shin:Aligning a parallel Korean-English corpus at the word and phrase level,NLPRS'95,1995
[2] Kuang-hua Chen;Hsin-hsi Chen:A part-of-speech-based alignment algorithm,COLING'94,1994
[3] Pascale Fung; Kenneth Ward Church: 1994 K-vec A new approach for aligning parallel texts, COLING'94,1994
[4] Kenneth Ward Church:Char-align:a program for aligning parallel texts at the character level·31st Annual Meeting of the Association for Computational Linguistics,1993
[5] Peter F·Brown,et al·:Aligning sentences in parallel corpora,Proceedings of 29th Annual Meeting of ACL,1993
[6] William A·Gale,et al:Extracting noun phrases from large-scale texts:a hybrid approach and its automatic evaluation·Proceedings of 32th Annual Meeting of ACL,1994
[7] 赵铁军、李生 等:双语语料库研究中的若干统计结果,《计算语言学进展与应用》,1995
[8] 解建和 等:从大规模语料库看双语词典释义的适用性和客观性原则,《计算语言学进展与应用》,1995

(责任编辑 许增棋)

《情报学报》1997年增刊通知

《情报学报》是国家一级学术刊物,在图书情报领域享有盛誉。由于受版面的限制,许多作者投至本刊的稿件未能在本刊上发表。为满足广大作者的要求,经研究,决定在1997年下半年出版一期增刊。

一、征文范围

情报学基础理论、信息加工处理的理论与技术、信息系统的网络化、电子出版物与电子图书馆、图书馆自动化技术、信息服务、信息产业、信息市场、用户研究、信息工作的组织管理、信息政策与立法等。

二、征文办法

由于《情报学报》的经费紧张,今年的增刊将采取收取版面费的办法,以贴补一部分出版费用。

三、稿件要求

1. 文章论点明确,主题突出。字数在3000字以上。来稿须附150-300字的中、英文摘要。

2. 文章按如下顺序叙述:(1)题目,(2)作者名,(3)作者单位名、所在地、邮编,(4)中文摘要,(5)中文关键词,(6)英文题目,(7)英文作者名,(8)英文作者单位名、所在地、邮编,(9)英文摘要,(10)英文关键词,(11)正文,(12)参考文献。

3. 来稿请附50字左右的作者简介,包括:姓名,性别,出生,职务/职称,所从事的专业领域或研究方向,主要成果。

4. 务请在稿件左上角注明"增刊"字样。

5. 来稿一律不退。经审稿决定录用的稿件即发录用通知,并告之版面费数额。作者在2个月内未接到录用通知,稿件可自行处理。请勿一稿两投。来稿截止日期以邮戳日期为准,定为1997年8月30日。

来稿请寄:北京复兴路15号中国科技信息研究所《情报学报》编辑部,邮编:100038

联系电话:(010)68515544转2580。**传真**:(010)68514025