基于双语对齐语料－－英汉词典的自动生成

中文信息处理Project：0124120 杜浩

January 16, 2005

摘要 本文提出一种自动生成英汉词典的方法。该方法基于已对齐的双语语料库，针对查询英文单词，枚举其可能的中文释义，从中挖掘与该查询英文单词共生频率高，但自身衍生频率低的中文释义，生成词典。本方法在《圣经恢复本－中英对照版》语料库上实验，成功提取了5668单词的中文释义，并达到了1级释义正确率60% 和4级释义的正确率88%的评估结果。

1 引言

英汉词典，作为沟通中文和英语两种不同语言之间的桥梁，为中国与世界的交流做出了巨大贡献。它的编制和生产已经产品化。英汉词典编制的基础是中英文对于同一事物的两种不同表达。设想世界上第一本英汉词典的编制，由一个即懂中文又懂英文，但不知道中英文单词对应关系的人，经过长期对两种语言的学习、对比，他总结出对应的规律，发现在表达有关铅笔的时候，例如“There is a pencil on the desk. ”、“I have a pencil.”、“The pencil is beautiful.”，都含有“pencil”这一词。于是得出结论，“pencil”很可能指得就是“铅笔”。于是，将这一对应关系写下来，作为词典中的一个词条。

无论是印刷版词典，或是当今广为使用的类似“金山词霸”这样的电子词典，目前为止都是人工编制的。人经过对中英文的对比、总结，慢慢的发现词条之间的对应关系。人工智能领域研究和发展的今天，我们可以尝试用计算机来做这一件事。这可以看作机器翻译[1]的逆过程（一些早期的机器翻译算法，根据词典生成译文。而本文是根据双语对照的译文，生成词典）。这一工作的基础，是大量的双语对齐语料库，用来给计算机“学习”，恰好类似于第一个编英汉词典的人也需要学习大量对于同一事

物的双语表达。

引文[2]提出一种实现方案，在其实现步骤中，需要完成中文分词、释义词典过滤。然而，中文分词需要用到汉语词典，与此同时并不是语言中所有的词出现在词典中，音译英文姓名、地名往往不存在汉语词典中，分词的结果会使这些词的翻译不准确。另一方面，我们不妨假定词典生成这一工作从头做起，而不使用部分已释义的“释义词典”。

本文给出一个行之有效的统计方法，自动生成英法词典。此方法仅依赖对齐的双语语库，而不采用其它资源，在《圣经恢复本－中英文对照版》上取得良好的实验结果。

2 英汉词典自动生成方法

在这一节里，提出一种基于中英文对齐语料库的英汉词典自动生成的方法。基于这样一个原始的思想：给出英文单词，在双语语料库中查出所有包含该词的句，从这些句的中文翻译中，寻找高频出现的公共词语，另一方面，从中取出在整个语料库中低频出现的，这些词语与该英文单词共生共灭，它们有较大的概率成为该单词的中文释义。

2.1 给定英文单词，挖掘中文翻译

在叙述这一过程之前，我们假定对齐的双语语料库已采集好，共有n句，用集合$\mathcal{M}$表示，

$$\mathcal{M} = \{(E_1, C_1), (E_2, C_2), ..., (E_n, C_n)\} \tag{1}$$

其中(E_i, C_i)表示一句英文E_i对应它的中文句C_i。

给定查询单词w，第一步操作是创建候选释义集。例如给出查询单词“China”，认为它的可能的中文释义“中国”一定出现在包含“China”的那些句子的中文翻译里，于是，先采集出这些句。令$\mathcal{W}$是英文句中包含w的句集合，

$$\mathcal{W} = \{(E, C) | w \in E \wedge (E, C) \in \mathcal{M}\} \tag{2}$$

显然，

$$\mathcal{W} \in \mathcal{M} \tag{3}$$

下一步是从候选句中枚举所有的候选中文词。如果一句中文包含m个汉字，如果认为中文词可以任意长，则可以从这一句中搜取出$m(m+1)/2$个不同位置和长度的子串。由于$O(m^2)$的中文词量，数据量过大，另一方面，根据我们平常的经验，在词典中出现的常用中文词长度不会太长，于是可以做一点假设，设定阈值k为候选中文词的最大长度。给定一个中文句，将长度小于等于k的中文词枚举出来，做进一步考虑。

记$f_k(C)$为枚举单句中文释义操作。该操作将一个中文句C中一切长度分别为1, 2, 3, ..., k的中文子串截取，滤除包含标点符号的子串。那么，对于英文单词w，其所有可能中文释义集合$\mathcal{F}_w$定义为：

$$\mathcal{F}_w = \bigcup_{(E,C)\in\mathcal{W}} f_k(C) \tag{4}$$

怎样从$\mathcal{F}_w$中选出w的最可能的释义呢？这里采取的方法是对每个词进行评分。对于给定语料库$\mathcal{M}$、给定英文查询词w、给定的候选中文释义c，定义：

- 基频数n_w：查询英文单词w在$\mathcal{M}$的出现次数，基频数用来作为评分的参考。
- 共生频数$n_{w,c}$：候选中文释义c在$\mathcal{W}$的出现次数。
- 衍生频数n_c：候选中文释义c在$\mathcal{M}$的出现次数。

分别定义共生概率$p(c)$，衍生概率$q(c)$：

$$p(c) = \frac{n_{w,c}}{n_w} \tag{5}$$

$$q(c) = \frac{n_c}{n_{w,c}} \tag{6}$$

如上分析可知，p值高表明出现w的句也出现c的概率大；q值低表明不出现w的句出现c的概率小。评分的动机是，p高且q低的单词，应该打高分。下面举一例，可以帮助体会这一点：

一个对齐语料库$\mathcal{M}$（n较大），如果查询词w是“pencil”，经过搜寻，所有包含“pencil”的句，也即$\mathcal{W}$集合，在如下列出（共3句）：

- （There is a pencil on the desk，桌子上面有一只铅笔）
- （I have a pencil，我有一只铅笔）

- （The pencil is beautiful，这只铅笔很漂亮）

针对英语查询pencil，基频数$n_{pencil}=3$。作为举例，考虑候选中文词的其中四个：“桌”、“只”、“笔”、“铅笔”。共生概率计算得：p(桌) $=1/3$, p(只) $=1$, p(笔) $=1$, p(铅笔) $=1$。由于“桌”的共生概率很低，不大可能是pencil的中文释义。其余三个词，“只”，在其它句中有可能出现“一只猫”、“一只狗”，于是q(只)会比较大，同样，q(笔)较大；而q(铅笔)几乎能维持在1，因为一个句子中，英文不出现pencil而中文出现“铅笔”几乎没有可能。如此，“铅笔”的共生概率大，衍生概率小，“铅笔”在这四个候选中译词中，更可能是pencil的释义。这正是我们所期望的。

于是，采用一个简单有效的评分函数$S(c)$

$$S(c)=\frac{p(c)}{q(c)}=\frac{n_{w,c}^2}{n_c\cdot n_w} \tag{7}$$

将$\forall c\in\mathcal{F}_w$，分别应用于评分函数$S(c)$，得分高者，成为中文释义。

2.2 候选释义集的规模及其动态控制

上述过程直接操作，时间复杂度相当高，给定w，候选释义词的采样次数：

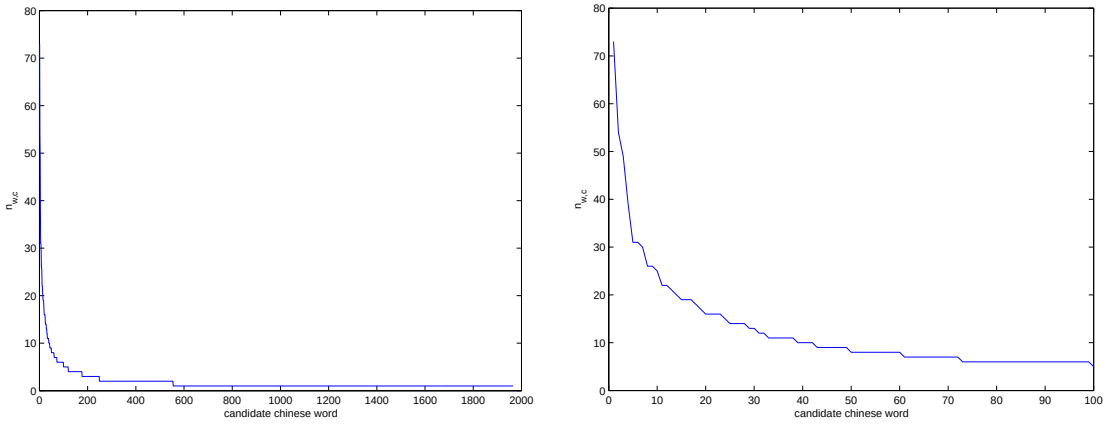
$$t=\sum_{j=1}^{n}\sum_{i=1}^{k}(|C_j|-i+1) \tag{8}$$

其中$|C_j|$表示中文句C_j的句长，对一个中等规模的对齐语料，$n=10000$, 取$k=30$，假定$avg(|C_j|)$=30，则$t=9,000,000$，对如此大规模语料做索引、排序、Hash都是相当大的开销。事实上，可以从两个方面可以大大减轻时间空间开销，而几乎不影响性能：

一、中文常用词，词长不超过4，不妨取k=4。

二、对于共生频率很低的中文候选词，可以在早期移除。

对于一个给定的英文词，大量的无关候选词的共生频数很低，这些词在早期，枚举出来之后不久，就可以移除。例如：w = “plant”，基频数$n_w=47$，将它的1966个中文候选词，按共生频数$n_{w,c}$排序，共生频数分布如图1(a)所示，可见，只有少数的词的共生频数较高。图1(b)是图1(a)前100个中文候选词的放大图。前56个词，及其共生频数如表1。发现，plant 的真实释义“载”、“种”、“栽种”的频数分别是39、26、20，都排在前20名。而共生频数较小的例如$n_{w,c}=8$的那些词，“果”、



(a) (b)

Figure 1: .

“果子”、“和”、“居”几乎与plant毫不相干了，以致排在后面1000多名的共生频数为1的那些候选词，可以忽略不计。

统计规律发现，所有英文词对应的候选中文词，其中八成的中文候选词，共生频数都是1、2，而高共生频数的候选词相当少。如果假定一个英文单词有两个不同的主要释义，例如，plant有“栽”和“种”两个意思，如果plant有$n_w = 47$次出现，其中有30将翻译成为“栽”，17次翻译成为“种”，则对于那些共生频数$n_{w,c} < 17$的中文候选词都可以删除，其中也包括那些为数八成的共生频数是1、2的候选词。

于是，动态删减候选词：将前述的“枚举中文候选释义”和“累计共生频数”两个操作同时进行，即，枚举$\mathcal{F}_w$的过程中，同步统计集合中每一个成员的共生频数。设定r_{max}、r_{normal}分别是中文候选词的总量上界、常界。一旦$|\mathcal{F}_w| \geq r_{max}$时，将$\mathcal{F}_w$的成员按共生频数排序，顺序移除共生频数最低的词直至$|\mathcal{F}_w| = r_{normal}$，继续枚举统计操作。

Table 1: 英文词“plant”的高共生频数的中文候选词

的	73	要	26	园	19	萄园	14	出	11	这	10	果子	9
们	54	不	25	必	18	葡萄园	14	其	11	中	10	耶	8
你	49	子	22	为	17	有	14	栽植	11	其中	9	耶和华	8
栽	39	所	22	葡	16	是	14	造	11	自	9	屋	8
我	31	你们	21	住	16	树	13	上	11	房	9	和华	8
他	31	栽种	20	萄	16	建	13	也	11	华	9	居	8
他们	30	地	19	葡萄	16	植	12	和	10	以	9	房屋	8
种	26	在	19	吃	15	得	12	并	10	果	9	就	8

Table 2: 英文词“hot”的高共生频数的中文候选词

的	39	就	9	日	7	有	6	时候	5	头	4	姓	4
人	15	在	9	那	7	为	6	使	5	炭	4	看	4
们	13	说	9	要	6	来	6	候	5	和	4	雅	3
我	13	这	9	都	6	对	5	下	5	便	4	之	3
他	12	了	9	饼	6	也	5	亚	5	去	4	处	3
热	11	是	8	上	6	出	5	从	5	百	4	样	3
你	11	他们	8	不	6	里	5	冷	5	耶	4	到	3
时	10	一	7	火	6	的时	5	的时候	5	百姓	4	出去	3

2.3 词典生成

在最终词典生成时，对于那些基频数很低的英文词，不应写入词典，这些词容易翻译出错。如前文所述，任意一个英文词，八成的候选词的共生频数$n_{w,c}$都是1或2，如果该英文词本身的基频数n_w就是1或2，无法从共生频率体现出，这些大量候选词究竟哪一个可能是真实释义。此时共生频率$p(c)$，参与在评分函数$S(c)$中几乎起不到作用。

设定一基频数下界L，称$n_w >= L$的英文词为有效词，作为词典元素，将其它词略去。按照对每个有效词c的候选中文释义的评分$S(c)$，取中文释义的高分者，生成词典。在词典中，附加释义和释义的评分值，评分值一定程度上能反映同一英文单词各个释义的使用频率。

3 实验及结果

本节为上文提出方法在真实语料库上的实测结果，包括语料库选取、参数设定、抽样评估、生成词典评估四个部分。

3.1 语料库选取

实验所用的中英双语对齐的语料库是《圣经恢复本》[3]。该语料库规模较大，中英对齐的互译句共有31090句。中英文对齐良好，翻译质量较高，翻译风格前后统一，语料库示意句如下：

- .....

- 第99句：And Lamech took two wives for himself: The name of the first was Adah, and the name of the second Zillah.
  第99句：拉麦娶了两个妻子，一个名叫亚大，另一个名叫洗拉。

- 第18910句：And they will build houses and inhabit them, And they will plant vineyards and eat their fruit.
  第18910句：他们要建造房屋，居住其中；栽种葡萄园，吃其中的果子。

- .....

下面是语料库准备的细节，流程图如图2所示。首先下载圣经恢复本双语对照版，是微软.chm电子书格式，该格式可以视为一个整合的网站，需要折分。“章节折分”操作将一个.chm文件，折分成可阅读的标准网页.html格式，这一步使用chm2web，结果将每一页分为一个独立的.htm，每个之中大约包含50句对照。接下来“句对抽取”自行分析.htm格式，得出标准文本文件。

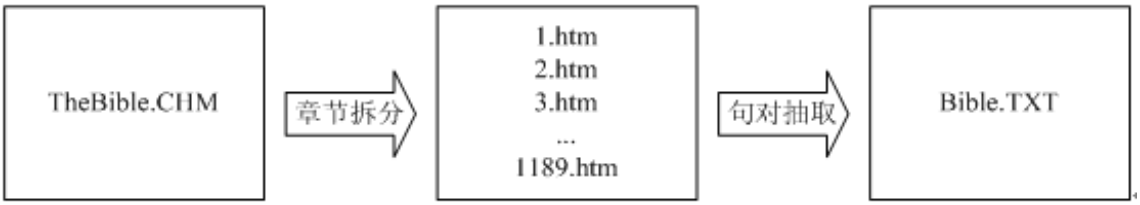


Figure 2: .

另外，电影的中英文对照字幕，也可作为可用的大规模双语对齐语料库。

3.2 参数设定

观察到中文词常用词以二、三、四字词为多，而此《圣经恢复本》中，成语使用不多，基中常用词以二、三字居多。设k=3。

候选中文词上界$r_{max} = 5000$和常界$r_{normal} = 3000$，以保证在枚举候选词过程中，r_{normal}覆盖有效候选词，使其不因为共生概率而剔除，另外时间空间又可以接受。

词典生成过程中，基频数下界$L = 5$，成功加入自动生成的词典中的英文词有5668词，每个词取得分最高的4个词义及其得分写入词典文件Dict.TXT。

3.3 抽样翻译评估

实验过程中，对给定的英文单词，所采用的评分函数$S(c)$对正确词义的区分度如何呢？随机抽取1个单词peter，得分较高的候选中文释义如图3所示： 由此可见

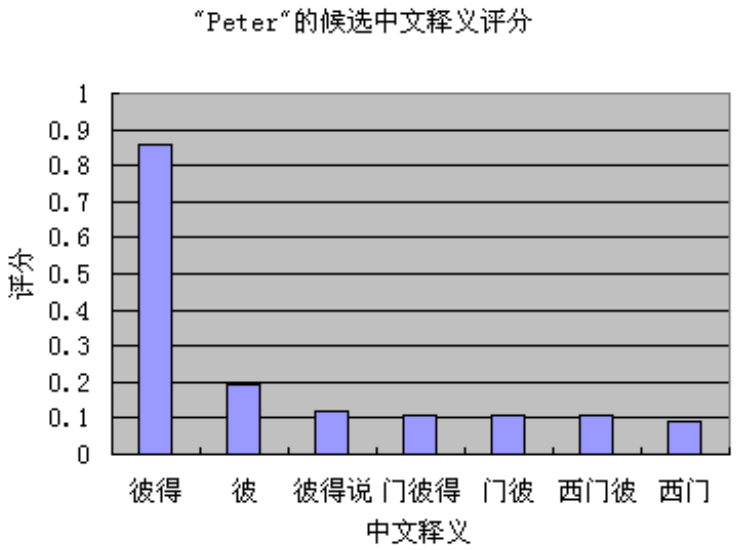


Figure 3: .

对于“Peter”一词，正确词义“彼得”得分最高，其它词义得分与它有一定差距。对于“Peter”的自动翻译准确。

同样，随机抽取其它词，也可做出类似的评估，这里随机列举一些，详请参见附件生成词典文件Dict.TXT.

Table 3: 生成词典随机抽样

affectionately		army		birth		prudent	
热切的	0.5	军	0.214638	出生	0.132031	精明	0.425676
切的和	0.375	的军	0.148514	出生地	0.125	精明的	0.321429
即到耶	0.25	军队	0.129253	生地	0.125	精明人	0.228571
的和祂	0.25	军兵	0.123862	生产	0.105263	精	0.15

3.4 生成词典评估

对本文所述自动生成词典方法在《圣经恢复本》中英双语对齐语料上的实验性能做评估。人工评估有一定的主观性，在此对本文评判符合“中文释义中包含其真正正

确词义”的标准做一说明：译文不会被误解，则认为正确翻译。例如，认为正确翻译的：

- wither 枯
- wither 枯干
- wither 枯干了
- voice 声
- voice 声音
- voice 的声音

认为不正确翻译的：

- under 日光之；（词义毫不相干）
- wither 河枯；（多了名词“词”的意思）
- thunders 听见雷；（多了动词“听”的意思）

由《圣经恢复本》生成的词典含英文词5668词，每个词按照评分取高分的四个词义。定义“自动生成的词典的r级翻译正确率”：

- 任取一个英文单词，其得分最高的r个中文释义中包含其真正正确词义的概率。

人工随机抽样50词，得1-4级正确率如表4所示：

Table 4: 《圣经恢复本》自动生成词典的1-4级正确率评估

一级正确率	60%
二级正确率	71%
三级正确率	76%
四级正确率	88%

4 体会

非常欣慰的做了这个设想和实验，一个简单的模型就能从双语对照语料中挖出一个大概有80%正确率的，还马马虎虎凑合着能用的英汉词典来！我想如果再仔细设计其它的评分方案，正确率和生成词典的可读性还能有提高的余地。

在实验本方法的同时，还有一个发现，写在这里，与大家共享：

《圣经》的单词量也少

《圣经恢复本》，共有英文单词782929个，但不同的词只有13931个。统计意义上，初中学生也能读懂圣经的87%，而GRE的单词量，足够了。参见图4。

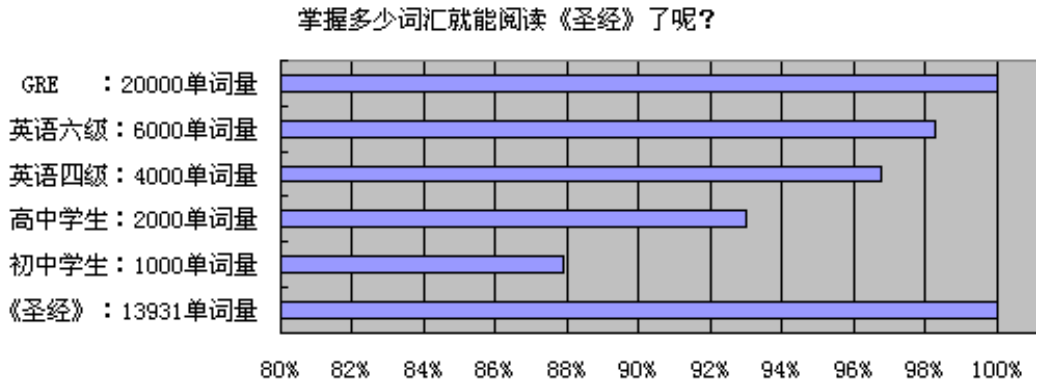


Figure 4: .

参考文献

[1] PF Brown, J Cocke, SA Della Pietra, VJ Della ... , *A statistical approach to machine translation*, Computational Linguistics, 1990

[2] 陈博兴、杜利民，《基于双语对齐口语语料的翻译词典的自动生成》，计算机学报，2003年03期

[3] 《圣经恢复本－中英文对照》，http://www.71630.com/bible/ch-en/index.htm