迭代策略和词典相结合的机器翻译词典获取*

刘晓月　杨沐昀　赵铁军

（哈尔滨工业大学计算机系，哈尔滨 150001）

E-mail: lxy, ymy, tjzhao@mtlab.hit.edu.cn

摘要： 翻译词典对于跨语言信息检索、计算机翻译等许多领域具有重要意义。为了解决基于汉英双语语料库的翻译词典获取问题，本文首先比较了四种常见的基于共现信息的词汇对译关系计算模型，并以对数相似性模型为基础，设计了一种迭代策略和词典相结合的汉英机器翻译词典自动获取的方法。初步实验表明，该方法的确能够提高翻译词典获取的正确率和召回率。

关键词: 翻译词典，汉英双语语料库，对数相似性模型，机器翻译

引言

由于翻译词典在全自动机器翻译[1][2]、各种辅助翻译工具[3][4]、跨语言信息检索[5][6]等众多自然语言处理领域具有重要的作用，所以利用双语语料库自动（或辅助）获取翻译词典一直是双语语料对齐研究中的一个热点，吸引了众多研究者的关注。

目前，基于双语语料库句子对齐结果的翻译词典自动获取技术大致可以分为以下几类：

1）利用 EM 迭代过程的无指导的双语词汇对译关系获取方法；

2）采用词汇对齐技术，通过双语词汇对齐结果获取翻译词典；

3）利用共现信息计算双语词汇之间的关联强度，从而建立词汇对译关系；

考虑到第一种方法计算量太大难以用于大规模双语语料；而第二种方法会受到汉英双语词汇对齐水平的限制，所以本文考虑尝试第三种策略，利用统计方法计算双语词汇之间的对译关系。在分析比较了几种常用计算词汇共现信息的统计计算模型的基础上，本文提出了一种迭代和双语词典相结合的方法用于汉英翻译词典的自动获取。实验表明，该方法可行性很高，不仅提高了模型精度，而且最大限度地利用了输入语料（高覆盖率）。

1 基于同现信息的计算模型

基于共现信息的双语词汇对译关系计算的基本思想是：双语句对的基础上，统计双语词汇的共现频率，进而计算出任意两个词对的关联强度（对译强度）。以汉英双语为例，如果某一汉语词 Wc 出现在一个句对的汉语句子中，则其译文 We 就会在该句对的英文句子中出现，若句对规模足够大，则象 Wc、We 这种互译词对的共现特征就会突显出来。根据共现频率计算任意两词对关联度的数学模型有很多种，其

* 本文研究受到国家 863 计划资助(项目编号：2001AA114101)。

中 Dice 系数，互信息，联列表[7]和对数相似公式[8]是其中比较常用的 4 种。

设已得到句子级对齐的汉英双语文本，其中源语（汉语）文本 C 中的词为 Wc，目标语（英语）文本 E 中词为 We。令：

freq(Wc, We) = (句对中 Wc 和 We 的共现次数)；

freq(Wc) = (Wc 出现总次数)；

freq(We) = (We 出现总次数)；

N = (全部句对数)；

则：

Dice 系数(Dice Coefficient)：

$$h_{DICE}(Wc, We) = \frac{2 \times freq(Wc, We)}{freq(Wc) + freq(We)};$$

互信息(Mutual Information)：$h_{MI}(Wc, We) = \log \frac{P(Wc, We)}{P(Wc) \times P(We)}$；

联列表方法(Contingency Table)：$h_{CT}(Wc, We) = \phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$；

其中：

a =freq(Wc, We)，b =freq(We)-freq(Wc, We)，c =freq(Wc)-freq(Wc, We)，d=N-a-b-c。

对数相似公式 (Log Likelihood)：

$$H(Wc, We) = 2[\log L(p1, a, a+b) + \log L(p2, c, c+d) - \log L(p, a, a+b) - \log L(p, c, c+d)]$$

其中：

$\log L(p, n, k) = k\log(p) + (n-k)\log(1-p)$，$p_1 = a/(a+b)$，$p_2 = c/(c+d)$，$p = (a+c)/(a+b+c+d)$，而 a、b、c 和 d 的意义同联列表方法(Contingency Table)中的 a、b、c 和 d 的意义。

一般来说，Dice 系数的值介于[0, 1]之间。数值越大，表二者共现频率越大，越有可能成为对译词汇。互信息公式是一种基于信息论中的互信息概念来计算词间关联程度的方法。联列表模型是 Gale 和 Church 等人设计的一个服从χ^2分布的随机变量。而对数相似性公式是 Dunning 根据二项式分布设计的判断任意词对的关联强度的公式。

2 模型的性能对比

为了对比四种方法的性能，用一个 30094 对句对、且经过句子对齐加工的汉英双语语料库进行实验。实验前，对其中的汉语句子进行了分词，对英语句子中的单词进行了词形还原。

在一个双语句对中，每个汉语词和对应的英语句子中每个单词视为一次共现，同时令 a = freq (Wc, We) = Min { freq (Wc), freq (We)}。这样，语料中出现的所有共现频率大于 1 的汉英词对共计 287664 个，其中汉语单词 10195 个，英语单词 6115 个。语料的基本数据如下所示（见表 1）：

表 1 汉英双语实验语料基本情况

	总词数 (含标点)	单词数	频度>1 的单词数	词对共现>1 包含单词数	句对总数
汉语	380,524 (已分词)	17,711	10,682	10195	30094

英语	324,302	10,688 (词形已还原)	6,581	6115	

本文的评价采用的是翻译专家独立于上下文来判别某个翻译词对是否合理。而考虑到词典编撰过程中效率因素，一般仅选择前 10 个候选作为有效译文候选，我们改进了文献[9]中的准确率计算方法：令首选译文位置加权系数为 1，次选译文位置加权系数为 0.9，以此类推，第 10 选译文的位置加权系数为 0.1。同时为了正确地反映不同译文候选的差异，我们进一步增加了正确结果的类型，一种是完全对等的译文，例如“动物-animal”，令其对译加权分值为 1 分；另一种是部分对应，即汉语词是英文单词译文的一部分或者英语候选译文仅仅是汉语译文的一部分，例如“美国-American”和“纽约-York”，它们的对译加权分值为 0.5。对于不正确的词对，令其对译加权值为 0，这样，每个汉语单词的总体加权得分就等于每个译文候选的加权得分（对译加权值×为值加权系数）的和。

考虑到汉英双语文本中分别有象“地、的”以及“the、a”等高频干扰词会产生噪声，为提高精度，采用了建立“停止词表（stop-word list）”的办法，把出现频率高于 1000 的英语词和汉语词列入表中，作为实验中剔除的高频干扰词。

经过这样的处理，全部汉英对译词对从 287664 个锐减为 150286 个，其中汉语单词为 8587 个，并不是简单的等于（10195-31）个。原因在于我们只考虑共现频率大于 1 的词对，在删除高频词时，使得和高频词共现的部分汉语和英语低频词也失去了候选资格。采用加权评价，计算全部四个模型的结果正确率（见表 2）。

表 2 删除高频词后四个模型的加权评价结果

	Dice	MI	CT	Log
完全正确译文	8557	8325	8560	8596
部分正确译文	2030	1983	2047	2061
加权得分	8571.65	8038.65	8591.05	8742.7
加权正确率	29.83%	27.97%	29.89%	30.42%
无正确译文的汉语词数	1357	1447	1359	1343

从表中数据我们可以看出，对数相似性公式的结果在四个模型中相对较好。该模型的全部汉语前十选译文中完全正确的有 8596 个，部分正确译文为 2061 个，都优于其余 3 个模型；而且其加权得分总计 8742.7，也是四个模型中最好的。所以我们可以认为：从翻译词典获取以及辅助词典编撰的角度来衡量，对数相似性公式是这几个模型中效果最好的。

3 迭代策略和词典相结合的双语词典自动获取

基于迭代策略和词典相结合的双语词典自动获取过程，步骤如下：

1°

2°

3°　　n 个汉英对译词对；

4°

5°　　4 时终止），重复步骤 2；

6°

7°

4 实验结果及分析

4.1 验证迭代方法是否有助于提高准确率

在删除高频词的对数相似性模型中选出前5000个词对作为模型的“直接结果”。随后采用上述控制策略，在同样条件下每次取前1000个，经过5次迭代，得到5000个汉英互译词对作为“迭代结果1”。最后，改变迭代步长，每次取前500个，迭代10次，这样的5000个英汉互译词对作为“迭代结果2”。三者的评价结果（见表3）：

表3 迭代策略的性能分析

Log模型（无高频词）	直接结果	迭代结果1	迭代结果2
含汉语词数	3298	3645	3713
完全正确译文	3412	3949	4049
部分正确译文	397	453	435
加权得分	3520.9	4061.65	4148.5
无正确译文的汉语词数	157	111	99

通过对比我们发现，迭代策略获得的结果明显提高，而且迭代步长越小，即每次从语料中删除的词对越少，性能越好。其原因类似“删除高频词”的道理：由于高频对译词对在语料中被删除，故语料中的噪声减少了，从而使计算结果的前几位候选变得更加可靠。

4.2 不同方法获取词典的性能比较

以每次选取500个汉英对译词对的方式进行了17次迭代，共获得8500个汉英对译词对。由于此时大多数汉语单词频率已经为4，对数模型的结果错误明显增加，所以使用一部现有英汉词典删除双语文本中所余的单词对译情况，共计9706个互译对。最后我们再次应用对数模型计算了文本中剩余的单词的互译强度。在所有这些词对中，如果最后的结果中包括高频词的话，那么这将是我们所能获得的最大词典，其性能见表4的“结果1”；如果不包括高频词，那么这将是我们获得的最可靠的词典，其性能见表4的“结果2”。

表4 词典获取性能对比

Log模型（无高频词）	结果1	结果2
互译词对总数	66696	32401
汉语词数	12749	11628
前10互译词对	36529	24315
完全正确译文	15383	15336
部分正确译文	1597	1481
加权得分	14956.35	14885.6
加权准确率	53.53%	72.50%
无正确译文的汉语词数	1737	668

从结果中我们可以看出，迭代策略非常有效的提高了传统模型的精度，而引入词典可以大幅度的提

高迭代的效果，模型的加权性能评价达到72.50%。特别值得注意的是，引入词典以后汉语词汇覆盖范围同时大幅度提高，最大可以达到（12749-1737）/17711≈0.62，已经获取了一部分词频为1的汉语单词的译文。

5 结论

本文探索了基于双语语料库的翻译词典获取的问题，比较了常用的4种基于共现的统计模型的性能，发现对数相似性模型是一比较优秀的计算同现信息的模型。在此基础上，使用迭代策略和词典相结合的方法获取双语词典，双语词典的引入极大提高了双语语料的利用率。

即便如此，这样的结果对于成功的编纂一部双语词典来说，还是不够的。分析错误产生的原因，一是间接共现的干扰，另一是因为汉英互译时其单词数目比实为一个m：n的问题，例："a lot of"翻译成"许多"就是一个3：1的典型范例，但目前我们只考虑了1：1的情况。

下一步工作，准备就以下几个方面进行改进：

1）迭代步长是否应随候选词对数的变化而做出相应的调整，通过某种机制计算出步长，再用参数校验测试其可信度。而不是从迭代1开始就一直采用相对不变的取步长策略；

2）迭代策略中汉英互译时单词数目比 m:n 的问题有待进一步解决。

参考文献：

[1] P. F. Brown, J. Cocke and S. A. Pietra et al. A Statistical Approach to Machine Translation. Computational Linguistics. 1990, 16(2):79-85

[2] Deryle Lonsdale, Eruko Mitamura, Eric Nyberg. Acquisition of Large Lexicons for Practical Knowledge-Based MT. Machine Translation, 9:3, 101-133, 1995

[3] I. Dagan, K. W. Church and W. A. Gale. Robust Bilingual Word Alignment for Machine Aided Translation. Proc. of Workshop on Very Large Corpora. 1993: 1-8

[4] Macklovitch. Using Bi-textual Alignment for Translation Validation: The TransCheck System. Proc. of the 1st Conference of the Association for Machine Translation in the Americas. Columbia, MD. 1994

[5] Nie Jianyun, Michel Simard et al. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining Parallel Texts from the Web. ACM-SIGIR Conference, Berkeley, California, 1999

[6] Jiang Chen, Nie Jian-Yun. Web Parallel Text Mining for Chinese-English Cross-Language Information Retrieval. International Conf. on Chinese Langauge Computing. Chicago, Illinois. 2000

[7] W. A. Gale, K. W. Church. Identifying Word Correspondences in Parallel Texts. Proceedings of the 4th DARPA Workshop on Speech and Natural Language. 1991: 152-157

[8] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19:61-74, March 1993

[9] Dekai Wu, Xuanyin Xia. Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. Machine Translation. 1995, 9(3): 135-163

作者简介：刘晓月（1978—），女，河北保定人，硕士生，主要研究方向为翻译模型和计算机辅助翻译技术；杨沐昀（1971—），男，讲师，研究方向机器翻译和计算语言学；赵铁军（1962—），男，博士，教授，博士生导师，研究方向是机器翻译和计算语言学。

Bilingual Corpus Based Chinese English Translation Dictionary Acquisition

LIU XiaoYue, YANG MuYun, Zhao TieJun

(Harbin Institute of Technology, Harbin 150001, China);

E-mail: lxy,ymy,tjzhao@mtlab.hit.edu.cn

Abstract: Translation Dictionary is significant to NLP researches like cross-language information retrieval, machine translation et al. In order to obtain a translation dictionary based on Chinese English bilingual corpus, this paper compares 4 models that are frequently used to measure word association. It further proposes a Chinese English machine translation dictionary acquisition method based on Log Likelihood model, iterative sized reduction and existing bilingual dictionary. Preliminary experiments indicate that this method can improved accuracy and recall.

Key words: translation dictionary; Chinese English bilingual corpus; Log Likelihood Model; machine translation