双语语料库研究中的若干统计结果

赵铁军　李　生　王海峰　杨　彦　毛永明　吴学伟

(哈尔滨工业大学计算机系，150001，哈尔滨)

Some Statistic Results Given By A Bilingual Corpus

Tiejun ZHAO, Sheng LI, Haifeng WANG, Yan YANG, Yongming MAO, Xuewei WU

(Harbin Institute of Technology, 150001, Harbin)

摘要：本文首先提出了以双语语料库作为机器翻译系统的质量标准。系统设计和调试时，让机器比照人工翻译结果，从而使机器译文达到新的水准。接着介绍了在英汉双语语料库基础上进行的两个统计工作。英语动词的词典译文和语料译文的匹配统计结果，表明要实现高质量的机器翻译，必须完成一部合适的机器词典。英汉句子长度对比的统计表明了以词为单位的英汉句子长度之间存在某种线性关系。

ABSTRACT: This paper firstly suggests bilingual corpora as the criterion of the MT systems. When an MTS is debugged, it refers to the results of manual translations so that the output of the MTS has achived a high profiency. In the following, the paper introduces two statistic results based-on an English-Chinese bilingual corpus: the matching of dictionary translations and corpus translations for English verbs and the comparing of the length of English and Chinese sentences. The former result shows a special machine dictionary is necessary for an idiomatic translation. The latter research shows that there exist some linear relations between the length of English and Chinese sentences counted by word..

1 双语语料库在 MT 中的应用

九十年代以来的语料库(corpus)语言学研究已经取得了可喜成果，语料加工技术(如标注)、基于语料库的机器翻译(如基于实例 EBMT 和基于统计 SBMT 的方法)、基于语料库的自然语言处理技术(如句法分析、词义消歧等)、知识抽取(如生成词汇表)、语料对齐等各个方面的研究都取得了进展。

双语语料库对 MT 的支持，无论它作为细粒度的知识源，还是作为 MT 测试的重要参考，目前都日益受到研究者的重视。如何利用双语语料库实现高质量的 MT，可以进行许多研究工作，如在语料加工基础上的 EBMT、从中抽取细粒度的语言知识等。这里我们想要讨论的是：把双语语料库当作一个翻译标准，让 MT 系统比照人工翻译结果进行系统设计和调试。其好处在于:

(1)帮助系统实现人员克服目标语语感不足的困难;

(2)主动适应用户对 MT 系统的苛刻要求;

(3)有利于发现和解决 MT 研究中遇到的精细语言现象。

我们目前在研的 BT863 汉英双向机器翻译系统采取了这一策略，并取得了初步成效。我们称其为面向实例的策略。在实践中我们认识到：双语语料库对于实现高质量的 MT 系统是非常必要的。它可以作为一个标准、一种方法、一种知识源来使用。

基于如上考虑，我们以双语语料库为基础进行了若干简单的统计工作。本文在第二节介绍了以语料库中的句子为标准对词典译文的检测，第三节介绍了获取双语句对的准备工作。所有结果都是在一个约有 2 万 1 千多句对的英汉对照语料库(以下简称为 ECP,English-Chinese Pairs)中统计得到的。该语料库容量约为 3MB，题材主要覆盖日常应用、商贸、一般科技等，具有一定的代表性。

2 英语动词的词典译文和语料译文的匹配统计

由于双语语料库是人工翻译结果，因此可以接受为地道的(idiomatic)译文。比照语料库中的实例进行翻译，MT系统的译文将不再局限于可懂(understandable)这样的主观标准，而把人工翻译结果视为客观标准。要想达到这样的目标，我们发现现有的词典资源不能满足要求。MT系统词汇级译文输出要依赖于词典，在译文变化较小的情况下，词典中的译文应该和语料句子中的译文相同或基本相同。根据这一前提，我们对BT863系统现有的英汉词典中的动词译文作了统计研究。

首先，简要介绍一下英语动词的汉语译文在语料和词典中出现的对比统计过程。

(1)从语料中取出一个句子，将其中的英语动词还原为原形，同时查词典，并把动词词组切分出来。凡是在动词词典中查到的词汇皆作为动词候选词。

(2)通过简单的上下文消除兼类。英语中动词与名词同形的情况也是比较常见的，通过查看该词前一个词的词性来排除不是动词的情况。如：Article+V/N -> ~V，从候选动词集合中删除该词。

(3)动词译文的匹配。将英语句子中每个动词的全部译文按顺序取出，到该句对应的汉语句子中查询。如果句子中包含该译文，则匹配次数加1。重复上述步骤到全部语料处理完毕。我们把动词的语料译文和词典译文相同的情况定义为译文一致，记为Tc=Td。

统计数据由表1给出。

表1　英语动词的有关统计结果

英汉词典中动词词条数	9563
语料句对中出现的动词数	3343
语料句对中出现的动词词次	33217
Tc=Td情况下的动词词次	11187

表中第一对数据说明语料中出现的动词数约是词典中动词数的 1/3，这表明有限的语料对词汇的覆盖程度总是有限的。假设我们的语料具有一定的代表性，则说明常用词的数

目是有限的。这和其他词频统计结果相类似。

表中第二对数据很能说明问题。前面我们已经说明，语料中的译文可以视作地道的译文，可是从词典中查出来的译文，只有 33 %能在语料中找到。余下的 Tc≠Td 情况占了多数。是不是词典中没有收录语料中出现的译文？通过进一步考察，我们发现这其中大多数的 Td 和 Tc 在意义上完全一样，就是词形不一样。因此机器无法匹配。

这一结果的明显解释是：单纯把一部书面词典输入计算机形成机器可读(machine-readable)词典，并不能很好地支持高质量的 MT。因为人在查词典进行翻译的时候，通常他不会把词典中的译文原封不动地写在句子中，而是要自己略加变通或修饰。这样才能使译文更流畅、更地道，而这对于机器来说却是极难做到的。所以，我们就不难理解高质量 MT 的困难了。如果没有一个面向 MT 的专用词典，我们又怎么能给出合适的译文呢！

由此可见，MT 系统的实现过程也是 MT 词典的完善过程，并且主要靠 MT 研究者自己来完成。为了帮助实现这一过程，我们设计了一个译文辅助获取程序。当 Td 不能在 Tc 中找到时，就利用一个对话框提示使用者输入合适的译文(见图 1)。在图 1 中，动词 return to 的几个译文均未出现在窗口所示的句子中，所以对话框提示可以在左上角的窗口中输入新的译文。

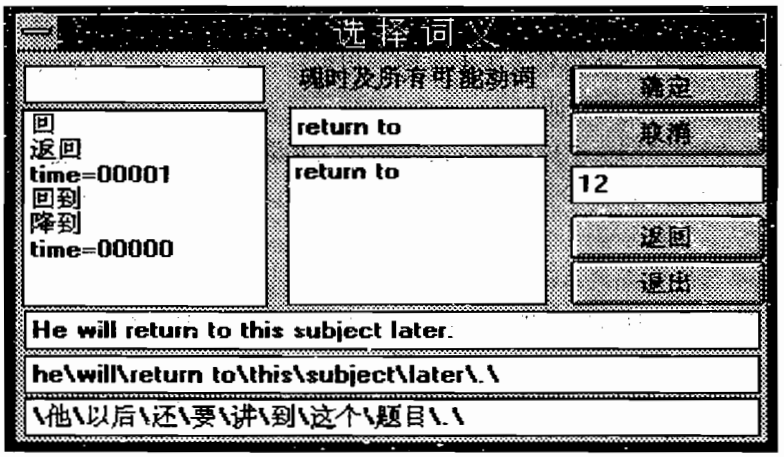

图 1　译文辅助获取窗口

3 英汉句子长度对比的统计结果

以双语文本(bilingual text)作为语料库来支持 MT 的研究，已成为当前国际 MT 界的一个研究热点。由于目前 MT 的基本翻译单位还主要停留在句子一级，因此人们把句对(sentence-pair)当作支持 MT 的基本单位。在利用双语文本的时候，搜集到的文本的并行程度是不一样的。可以是一本书的两种语言版本，文章级的对照本等。所以，从初始的双语文本到可进一步利用的双语句对语料库之间，还必须经过一个处理过程，这就是当前 CBMT

研究的一项重要内容——语料对齐(corpus alignment)技术。当然，除了句子级的对齐外，还可以进一步进行亚句级的对齐，那就需要更加复杂的处理了。

语料对齐技术的研究，已经提出了下述一些方法:

(1)根据两种语料的字节长度关系，已进行了英法、英日等语言的语料对齐研究[3]。

(2)根据两种语言词汇间的同源语关系，这对于许多西方语言语料之间的对齐是有用的[5]。因为它们的字母集基本相同，并且存在大量互借词汇(即所谓同源语)。但对于不同字母集合的语料间的对齐是不适用的，英汉语料就是这种情况。

(3)根据两种语料中词汇的分布关系。可用于生成双语词汇表[4]。

下面介绍的实验工作是根据 ECP 英汉双语句对语料库，进行英语句子和汉语句子长度间关系的统计研究。这一工作将为今后英汉句对的对齐提供某种依据。实验中以英语句子的长度为基准，英语句子长度和汉语句子长度的计算单位均取词数，对每个汉语句子首先进行了自动分词。

尽管文[3]提出英语和汉语句子的字节数间存在某种对应关系，但我们的实际统计表明，对于每个按字节计算的英语句长，汉语按字节计算的长度范围变化都很大，几乎无规律可言。因此我们舍弃了这种统计结果，长度单位采用了词数。由于翻译的基本单位是词对词的，所以以词数作统计单位是合理的。

图 2 - 图 4 显示了英语句长和汉语句长之间的统计分布。

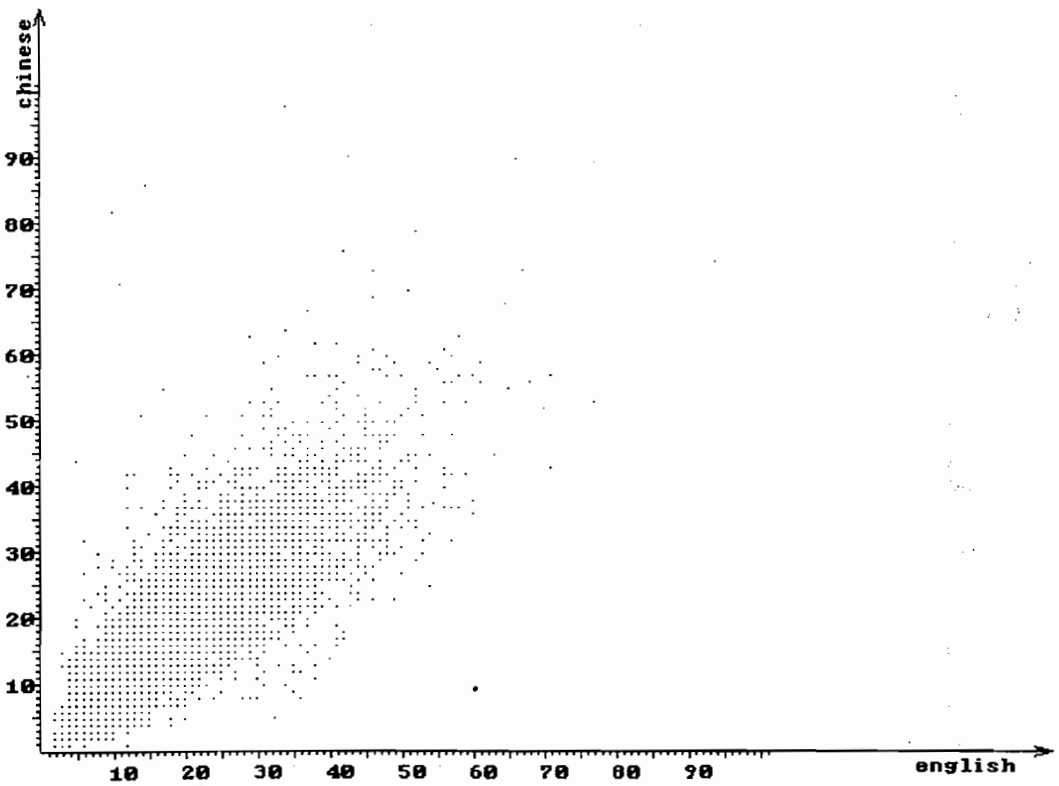

图 2　英语句长与汉语句长(词数)的统计分布

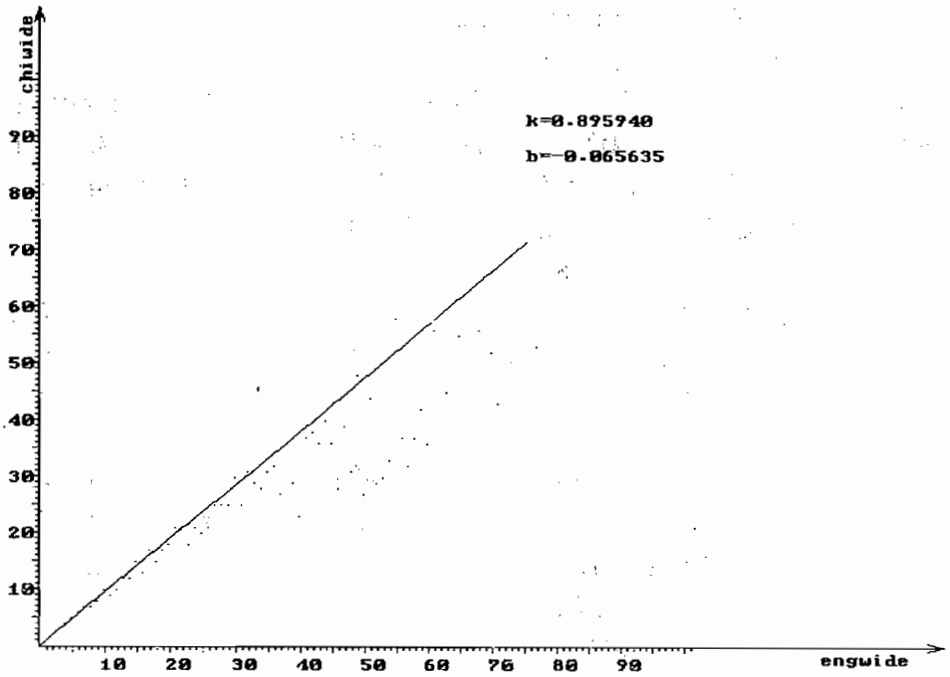

图 3　按最大概率计算的英汉句子长度(Le<30)的对应关系

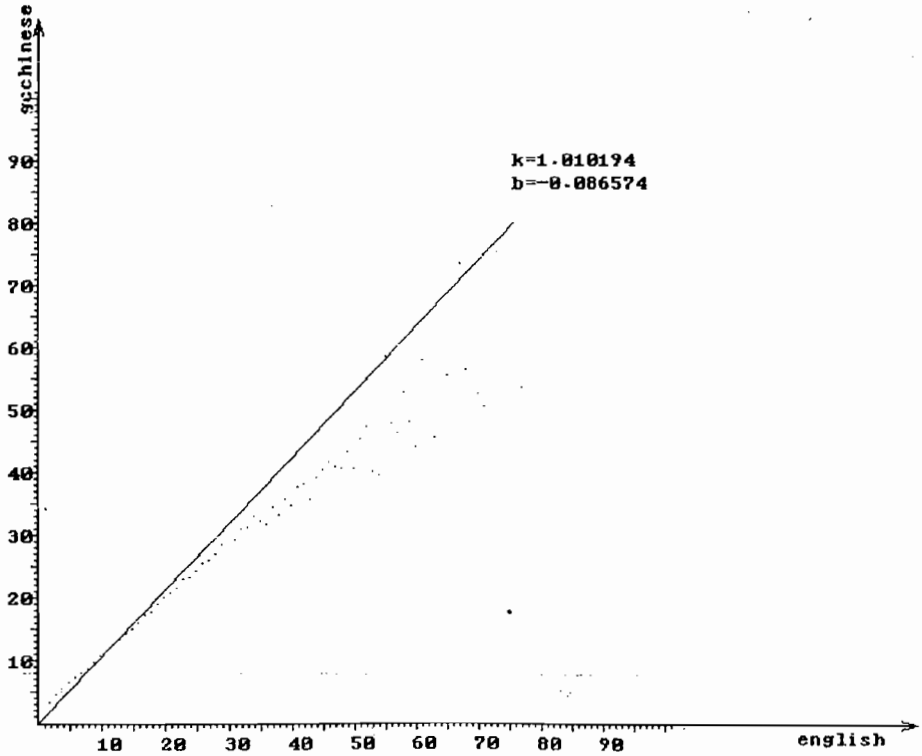

图 4　按数学期望计算的英汉句子长度(Le<30)的对应关系

在图 2 中，横纵坐标分别为英语和汉语句长(分别用 Le 和 Lc 表示)，句子的长度单位是词的个数。对一定句长的英语句子，其对应汉语句长可以看成是一个随机变量 X。我们可以把统计中出现的最多次数的汉语句子某个长度作为 X 的最大概率。在计算数学期望(均值)时，我们同样用汉语句长的统计频率来代替概率。图 3 和图 4 表示按最大概率和按数学期望计算的英汉句子长度之间的对应关系。在 ECP 语料库的全部 21539 句对中，英语句长 ⩾ 30 的句子只有 1015 个，因此这一部分的语料数量不足，在统计时没有考虑。余下的句长=1—29 的情况其样本空间(被统计的句子数)都较大，均为 200 以上。在用最小二乘法进行曲线拟合时，我们只使用了这部分数据。

图 3 和图 4 表明英汉句长间存在某种线性关系，因此用最小二乘法求出了所示直线。其回归方程为 y=kx+b。图中右上角给出了 k 和 b 的值。其中:

$$k=\frac{\sum_{i=1}^{29} x_i y_i - (\sum_{i=1}^{29} x_i \sum_{i=1}^{29} y_i)/n}{\sum_{i=1}^{29} x_i^2 - (\sum_{i=1}^{29} x_i)^2/n} \qquad (1)$$

$$b=\frac{\sum_{i=1}^{29} y_i - k\sum_{i=1}^{29} x_i}{n} \qquad (2)$$

在上式中，x_i表示英语句长 Le，y_i表示汉语句长 Lc，n 是计算时所取的点数(即 29)。从图中我们也可以看出当英语句长超过 30 时，和回归直线相差都较大。

当 Le 一定时，Lc 作为随机变量 X 服从某种概率分布。我们考虑 X 服从正态分布。尽管句子长度是离散型随机变量，但这里借用了连续型的正态分布概率密度来计算其分布。对于正态分布，其概率密度

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (3)$$

这里 x 值实际是离散的，取值范围是 0<x<100。由于正态分布的μ和σ分别是其数学期望 E(x)和方差 D(x)，所以我们只要求出 E(x)、D(x)，即可得到其概率密度函数。

$$E(x)=\sum x_i p_i = \sum x_i f_i = \sum x_i \frac{n_i}{n} \qquad (4)$$

$$D(x)=\sum [x_i - E(x)]^2 p_i = \sum [x_i - E(x)]^2 \frac{n_i}{n} \qquad (5)$$

在上两式中都用统计频率来近似地替代概率，其n_i表示对应于指定的英语句长而汉语句长为x_i的句子数，它和给定英语句长的全部句子数n之比就是x_i的频率。我们对英语句长<30 的汉语句长分布情况作了比较，只要语料充足(句子数>1000)，则结果基本符合正态分布。这里仅举一例说明。图 5 表示英语句长 = 10 的汉语句长的统计分布情况。此时，E(x)=11.067909,D(x)=3.218162，句子总数是 1664 个。图中的横坐标是汉语句长，纵坐标是对应于汉语句长的句子个数。孤立点表示实际统计结果，正态分布曲线是根据概率密度和句子总数的乘积结果而画出的。对于离散型变量来说，区间的概率都集中在端点上。这

样，将每个句长点代入概率密度函数就可得到该点的概率。它和句子总数的乘积，便相当于该句长点按正态分布的句子数。从图中我们可以看到，实际统计点的分布和正态曲线是相当吻合的。

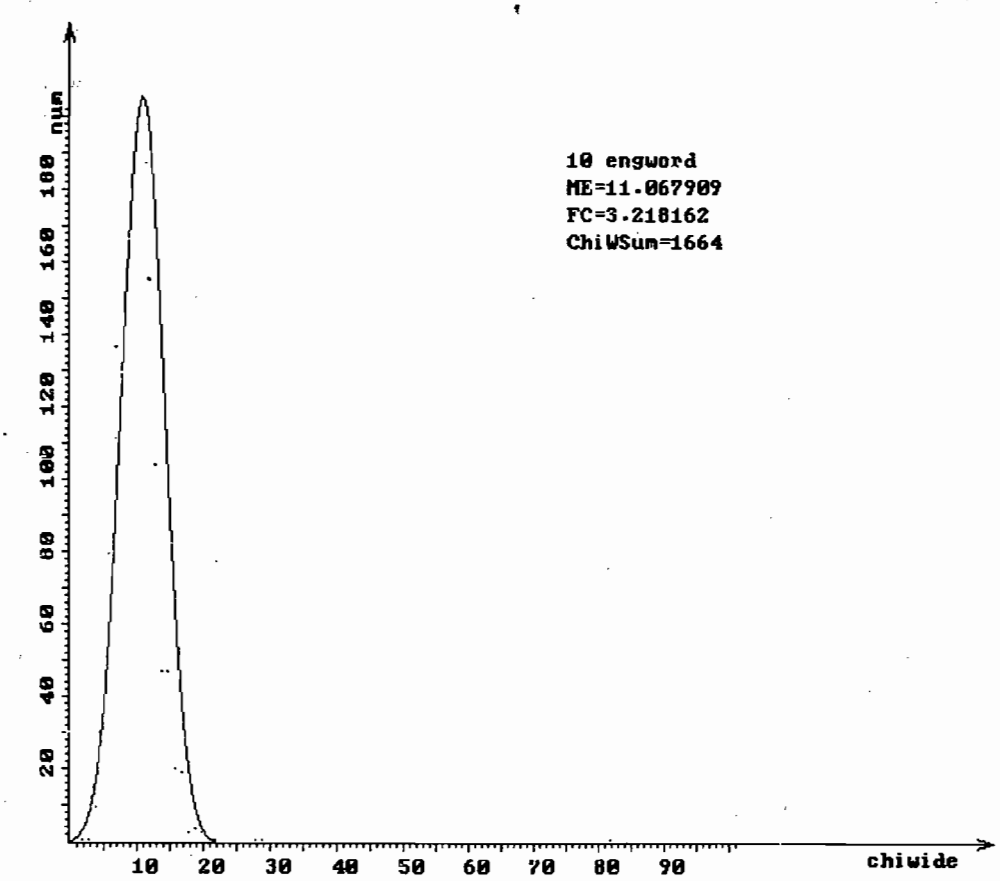

图 5　给定 Le(=10)条件下 Lc 出现次数的统计分布

参考文献

[1]张培荣主编，概率论与数理统计，东北财经大学出版社， 1993

[2]赵善中、黄春湛，计算方法，哈工大教材， 1980

[3]Church, K.W. ,Dagan,I. ,et al, Aligning Parallel Texts: Do Methods Developed for English-French Generalize to Asian Languages? Lecture in Qinghua University, 1994

[4]Fung, P. & Church, K.W., K-vec: A New Approach for Aligning Parallel Texts, COLING-94, pp1096-1101

[5]Simard,M. ,Foster, G., & Isabelle, P. , Using Cognates to Align Sentences in Bilingual Corpora, TMI-92, pp67-82