基于长度算法的中－英双语文本对齐的试验*

刘昕　周明　黄昌宁

清华大学计算机科学与技术系

北京　100084

摘要

双语语料库文本对齐的研究已成为新一代机器翻译中的一个至关重要的问题．已经发表的对齐算法对处理两种同属西方语系的语言取得了很好的效果．本文论证了基于长度的对齐算法适用于中－英文本的对齐，并依照此思想，设计并实现了一个文本对齐的原型系统．试验结果表明，该系统性能达到了令人满意的程度．

An Experiment to Align Chinese-English Parallel Text Using Length-Based Algorithm

Xin Liu　Ming Zhou　Changning Huang

Dept. of Computer Sicience & Technology, Tsinghua University

Beijing　100084

ABSTRACT

Bilingual texts alignment becomes a crucial issue in the new generation of MT. Many alignment algorithms have been proposed, with very high precision. But the languages they mostly processed both belong to occidental family. In this paper, we first illustrate the applicability of the length-based algorithm to the Chinese-English text. Then we design and implement a prototype of the alignment system. The experimental result shows encouraging perfermance.

一 引言

自八十年代以来，基于统计（Statistics-Based）和基于实例（Example-Based）方法的出现及广泛应用给机器翻译的研究工作注入了新的活力，标志着机器翻译进入了一个新时期（Hutchins,1993）．这两种方法共同的特点是：都需要一个双语语料库（Bilingual Corpora）直接或间接地作为翻译的知识库．这种双语语料库中包含了源文和译文相互对应的语言信息，是支持机器翻译的最为宝贵的资源．目前，双语语料库已经被广泛应用于自然语言处理的许多方面，如基于实例的机译研究、语义排歧、双语词典和术语库的建立、翻译知识的获取等，而且都取得了重大的进展．

*本项目得到了航天预研基金和国家自然科学基金的资助

双语语料库可以有很多种组织形式，其中最为常用的一种是：以句为基本单位，把源文和译文按照句子间的互译关系配对放在一起（对齐，alignment）。但是在绝大多数原始的双语文本中，都没有实现这种语句间的配对，因此要生成这种形式的双语语料库，就必须对原始文本进行对齐工作。如果对齐是通过人工的手段来进行的，其效率和可靠性都很差，所以利用计算机来自动进行双语文本的对齐，已成为一个亟需解决的问题。

在下一节中，我们先介绍一种基于长度的句子对齐算法；第三节首先论证该算法对于中－英文双语文本的对齐的可行性，然后描述了一个基于此算法的对齐试验，并对试验的结果加以分析；最后一节提出一些对算法的改进措施并讨论了本次研究工作的应用前景。

二 基于长度的对齐算法

对一个双语文本的对齐就是要判断源文中的每一个部分分别对应为译文中的哪一个部分。这里的一个“部分”可以是一个段落、一个或几个句子，一个词或是一个短语等等，它们分别对应着段落的对齐、语句级的对齐和词或短语的对齐等。本文只限于讨论语句级的对齐。

双语文本的自动对齐存在着许多困难。以句子级的对齐为例，由于在翻译过程中，每个句子并非都是一句翻译成一句，还存在大量的一句翻译成多句、或者多句翻译成一句，有时甚至还会有多句翻译成多句的情况。Gale 和 Church 曾经对一个英－法－德三语语料库 UBS 进行了这种句子翻译复杂性的统计，得出了如下的一个结果（表 1）。这种翻译过程中的多样性导致了文本的自动对齐难度很大。因此，如何寻找一种高效的双语文本自动对齐算法，也就成为国际机器翻译界的一个热点课题。

表1 句子间各种配对方式的频率(Gale & Church, 1991)

句子配对方式	出现频率
(1,0) 或 (0,1)	0.099
(1,1)	0.89
(1,2) 或 (2,1)	0.089
(2,2)	0.011

自 Kay 和 Roscheisen （1988）以来，围绕着对齐所进行的研究工作主要还有：Brown et al.（1991）、Gale 和 Church（1991）、Simard et al.（1992）以及 Chen（1993）等。总的来说，他们所提出的对齐算法基本上可以分成两类：基于长度的方法和基于词汇的方法（或者是二者结合）。第一种方法是以句子长度作为对齐的主要标准，通过统计的方法，寻求在这一标准下最优的句子级互译匹配模式。与之相对应，第二种方法是分析两种文本中各句子的词的信息，包括词性、构词法等，通过词的配对来寻求句子之间的配对。其中，第一种方法对资源的要求比后者要少，它基本上不需要其它辅助的机器词典，所以执行效率明显高于后者。我们所用的算法就是一种基于长度的算法。这种算法的基础是基于概率的理论，下面我们通过一个简单的模型来分析它。

考虑一个双语的语句对（Tc,Te），其中 Tc 是源文，长度为 m 个字节；Te 是译文，长为 n 字节。我们把它们按照源文的长度分为 m 个独立的部分离散地来看：

Tc = C1C2 … Cm，　　Ci 长为一个字节，i=1...m；

Te = E1E2 … Em，Ei 是 Ci 对应的译文，它可以是一个或多个字节，i=1...m。

则在长度标准下，语句对（ Tc,Te ）能够对齐的概率是一个条件概率 Prob （ Tc⇔Te|m,n ）。利用 Bayes 定理有:

$$Prob(Tc \Leftrightarrow Te|m,n) = \frac{Prob(m,n|Tc \Leftrightarrow Te) * Prob(Tc \Leftrightarrow Te)}{Prob(m,n)} \qquad (1)$$

对特定的 m 和 n 来说，上式中的分母项是一个常数，所以可以忽略．则公式（ 1 ）化为:

$$Prob(Tc \Leftrightarrow Te|m,n) = Prob(m,n|Tc \Leftrightarrow Te) * Prob(Tc \Leftrightarrow Te) \qquad (2)$$

第一项 Prob(m,n|Tc⇔Te)表示：在对齐情况下，长为 n 的译文以多大概率与长为 m 的源文相对应．如果我们把所有的 Ci⇔Ei 看成是一系列独立同分布的随机事件，并且设它们的分布具有期望 c 和方差σ^2，则由概率学中的 Liapunov 定理可知 Prob(m,n|Tc⇔Te)服从正态分布 N(c,m*σ^2)．这样我们可以用随机变量δ(m,n)来代表这一项，为使δ 服从标准正态分布，取:

$$\delta(m,n) = \frac{n - c * m}{\sqrt{m * \sigma^2}} \qquad (3)$$

这样，我们只要对特定的 m 和 n 求随机变量δ 的值，然后再计算对应的标准正态分布概率值就能得出 Prob(m,n|Tc⇔Te)。

第二项 Prob(Tc⇔Te)表示：在所有（不考虑长度）的情况下，文本 Tc 与 Te 配对的概率值，我们这里是用 Tc 与 Te 的句子配对模式的概率来衡量它。常见的配对模式概率可以由统计方法得到，如前面的表 1 中列出的就是一些常见配对模式概率的经验值。

以上所求的是双语文本段中的一个语句对对齐的概率值，我们通过取负对数将它化成两项求和:

$$\begin{aligned} Score((Tc,Te)) &= -\log(Porb(m,n|Tc \Leftrightarrow Te) * Prob(Tc \Leftrightarrow Te)) \\ &= -(\log Prob(m,n|Tc \Leftrightarrow Te) + \log Prob(Tc \Leftrightarrow Te)) \end{aligned} \qquad (4)$$

它的值是用来评价该语句对对齐的程度，所以我们把它称为一个语句对的评价值．而对整个双语段落对齐的评价值就是组成该段落的所有语句对评价值之和．要寻求一个双语段落中语句的最佳对齐方案，就是要求该段落对齐评价值的最小值．这里可以使用动态规划算法来求解．对于一个包含有 s 句源文和 t 句译文的双语文本段落来说，我们可以用如下的递归式来描述这种动态规划算法:

$$D(0:s,0:t) = \min_{(x,y)} (D(0:s-x,0:t-y) + Score((s-x+1:s, t-y+1:t))) \qquad (5)$$

其中 D(0:s,0:t)表示一个段落中的前 s 句源文与前 t 句译文对齐的整体评价值，并且（ x,y ）是一种可能出现的语句配对，如表 1 中的各种配对方式。

三 中－英文双语文本对齐的试验

虽然以上的算法在对两种西方语言文本的对齐中，取得了很大的成功（ Gale & Church, 1991 ），但是我们这里要试验的是中－英文的对齐，所以需要作进一步的考察．为了论证这一问题，我们人工对齐了一个各具有 17 个段的中－英双语文本，共得到了 171 个语句对，然

后对每个语句对中的中－英文长度进行统计，结果如图 1 所示。从图中，我们可以看出中文和英文长度的线性关系很强。

另外，为了进一步验证随机变量δ在中－英文下的分布情况，我们统计了δ的分布密度（未计入分母中的常量σ^2），其结果如图 2 所示。从该图中来看，我们可以认为δ近似服从正态分布。

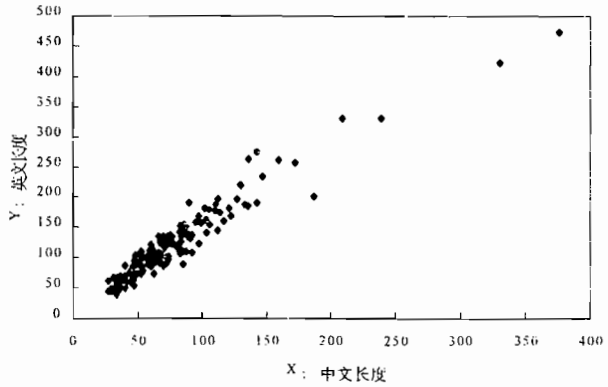

图 1　中－英文语句长度的关系

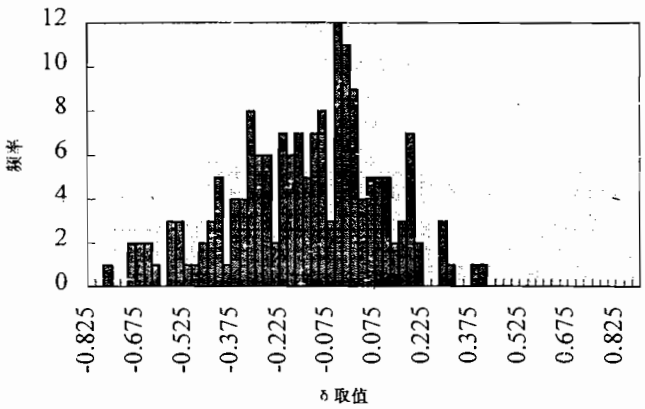

图 2　随机变量δ分布示意图

通过上面的分析，我们认为：这种基于长度的算法在中－英文的对齐中仍然适用．因此，下一步我们就要确定实施该算法所需要的参数，即各个独立分布的方差和期望，它们在也都是在上面的统计中得出的．其中的期望值是整个双语文本中的英文和中文长度的比值，实际得出：c=1.46；在公式(3)中，我们可以看出整个语句对的方差值与中文句子长度成正比，而σ^2正是该比值，所以通过统计每个语句对方差值与中文长度的比值，可实际得出：$\sigma^2 = 2.9$。

另外还需要考虑的一个问题是在中－英文中各种句子配对模式的概率值是否和 前面所述 Gale 的经验值相同．我们在对试验用的语料库统计中发现，没有出现(0,1)或(1,0)和(2,2)的情况，但是增加了(1,3)和(3,1)的情况．实际的统计结果如下：

表 2　中－英文句子配对模式统计

语句配对模式	出现频率
(1,1)	96.9%
(1-2) 或 (2,1)	2.7%
(1,3) 或 (3,1)	0.4%

我们使用的测试语料库是一份有关 DEC 表格处理工具的手册，它一共分为四个文档，每个文档包括相对应的两个中－英文文件，它们的具体情况如下表 3 所示．为了进行试验，我们对它进行了一系列预处理工作，即先把它们的段落作了对齐，然后对段落中每个句子（包括中、英文）的句尾边界进行了标记．此外，为了对对齐结果进行检验和分析，我们还对这些文本做了人工对齐．

表 3　测试用语料库规模

文档编号	段数	语句对数	c 统计值	σ^2 统计值
1	17	171	1.51	2.7
2	186	640	1.49	2.9
3	315	1012	1.47	3.1
4	603	2306	1.43	7.9

基于上面介绍的长度算法，我们采用 Visual C++ 编程实现了一个 Windows 环境下的对齐试验系统，该系统运行于一台具有 8M 内存的 486 DX2/66 的机器上，通过测试对齐速度达到了 110 句/秒．另外对上述四篇文档对齐结果的出错率统计如下:

表 4　对齐结果统计

文档编号	语句对个数	出错个数	错误率
1	171	0	0%
2	640	0	0%
3	1012	0	0%
4	2233	45	2%

在表 4 中我们可以看到错误全部集中在文档 4 中．分析其原因是因为文档 4 文本规模最大，并且　　中－英文句子长度关系的方差也最大．从而导致中－英文句子间的线性关系减弱，使对齐出现错误．

通过对错误的分析，我们发现出错的大部分都是下面这种情况：即，如果译文对源文的大部分内容都不作翻译，直接就照搬过来（这种情况在科技文章是屡见不鲜的），那么这两个句子的长度比就接近于 1，与系统中所用的 c=1.46 很有差距，此时如果下面紧接着的一个句子比较短，那么就会出现错误的对齐结果　（例 1），并且该错误极有可能会一直影响到段落的结束．这种错误的扩散性正是基于长度算法的最根本的缺点，下一步我们可以考虑在长度算法的基础在加入一些词汇的信息来对它进行改进．

例 1　对齐结果中的出错实例

… … …

〖中文〗 参数 COLOR-NAME 是 UNCHANGED、BLACK、BLUE、GREEN、CYAN、RED、MAGENTA、YELLOW、WHITE 之中一种。

〖英文〗 Parameters COLOR-NAME is one of: UNCHANGED, BLACK, BLUE, GREEN, CYAN, RED, MAGENTA, YELLOW, WHITE.
This can be specified as three RGB numbers.

… … …

四 结束语

本文论证了基于长度的对齐算法对中－英文本依然适用，并且通过实际的测试结果表明，使用该算法能够获得令人鼓舞的结果．但是就系统的实用性和稳定性而言，还有一些方面需要改进。

本文只使用了基于长度的信息，如果再利用有关词汇的信息，如同源词（指中、英文本中直接对照此）和关键词性（即名词趋向于译成名词等），则对齐正确率还会有所提高．另外在研究中发现，为进行对齐而进行的预处理工作花费了很多的工作量，如确定英文句尾边界等。下一步研究拟增加一些新的手段，减少这一部分的人工时间。

这次研究工作将服务于我们正在进行的基于实例的英－汉双向机器翻译课题。利用对齐手段，可以迅速获得大批英汉对照的例句库，它本身就可以在机器翻译支持系统中对译员提供帮助．另外，通过知识获取手段，可以进一步获取双语对译模板和双语词典，从而有效地提高基于实例机器翻译的性能，改善译文质量。

本研究得到了 AT&T Bell 实验室的 Church 教授的帮助，在此表示感谢。

参考文献

(Brown et al., 1991) Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. **Alignment Sentences in Parallel Corpora.** *Proceedings 29th Annual Meeting of the ACL,* pp. 169-176.

(Chen, 1993) Stanley F. Chen. **Aligning Sentences in Bilingual Corpora Using Lexical Information.** *Proceedings of the 31st Annual Meeting of the ACL,* pp. 9-16.

(Gale & Church, 1991) William A. Gale, Kenneth W. Church, **A Program for Aligning Sentences in Bilingual Corpora.** *Proceedings of the 29th Annual Meeting of the ACL,* pp. 177-184.

(Hutchins, 1993) Hutchins,J., **Latest Developments in machine translation Technology.** *MT Summit IV*, Kobe, Japan, July 1993

(Kay & Roschesein, 1988) Martin Kay, Martin Roscheisen, **Text-Translation Alignment.** *Computational Linguistics*, 1993, Vol. 19, pp. 121-142. (1988 年的原稿是一份 Xerox Palo Alto 研究中心的技术报告)

(Simard et al., 1992) Michel Simard, George F. Foster, and Pierre Isabelle, **Using Cognates to Align Sentences in Bilingual Corpora.** *TMI-92,* pp. 67-81.