(7) 40-44, 57.

中 文 信 息 学 报
第 14 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 14 No. 6

基于未对齐汉英双语库的翻译对抽取

TP391.1

王 斌
(中国科学院计算技术研究所软件研究室 北京 100080)

摘要:本文主要研究基于未对齐的汉英双语库翻译对抽取。文章首先介绍了 Pascale Fung 在这方面设计的两个算法。在此基础上,文章对后一种算法进行了部分的改进,使得其更适合于真实双语文本的翻译对抽取。实现结果表明改进后算法的有效性。本方法可以用于基于大规模双语语料库的短语翻译抽取、词典编纂等应用,具有较高的应用价值。

关键词:双语库;对齐;翻译对;抽取;自然语言处理 汉英双语库, 未对齐.
中图分类号:TP391

Translation Pairs Extraction from Unaligned Chinese-English Bilingual Corpora

WANG Bin
(Software Division, Institute of Computing Technology, Chinese Academy of Sciences Beijing 100080)
E-mail: wangbin@mtgroup.ict.ac.cn

Abstract: This paper focuses on extracting translation pairs from unaligned Chinese-English bilingual corpora. First, it introduces two methods proposed by Dr. Pascale Fung. Then, we revises the latter one to satisfy the need of real texts. The experiment results show the effectiveness of our method and it can be applied widely in many NLP applications such as phrase extraction, bilingual lexicography, etc.

Keywords: bilingual corpora; alignment; translation pair; extroution; NLP

一、引言

语料库语言学的兴起,使自然语言处理进入了一个崭新的阶段。大规模的语料库不仅可以为自然语言处理提供实例佐证,而且可以从中获取语言学知识加以利用。其中,由双语对照文本组成的双语语料库,更是由于其包含的双语间的翻译关系而受到自然语言工作者的青睐。

收稿时期:2000-05-08
基金项目:国家 973 二级子课题(G1998030510)
作者**王斌**,男,1972 年生,博士,副研究员,主要研究方向:自然语言处理,信息检索,知识挖掘.

目前，基于双语库的工作主要包含两方面的内容。一方面，人们把搜集到的双语生语料加工成熟语料；另一方面，人们可以从双语库中抽取各种知识加以利用。比如，人们可以从大规模对齐的双语语料库中抽取词翻译对、短语翻译对甚至翻译模板，而这些资源对于机器翻译、跨语言检索、词典编纂等自然语言应用都具有极其重要的使用价值。

所谓翻译对(Translation Pair，简称 TP)，是指互为翻译的源语言和目标语言片断构成的二元组。理论上说，这个片断可以是句子、甚至篇章，但是本文考虑的 TP 是词对或短语对。从双语语料库中抽取 TP，就是从双语语料库中抽取互为翻译的源语言和目标语言片断来。它根据所采用的双语语料库是否对齐分为两种：一是从已经做到句子甚至更细单位对齐的双语语料库中抽取 TP[4,5]；二是从未对齐(也可以说是全文对齐)的双语语料库中抽取 TP[1,2]。由于真正做到句子甚至更细对齐的双语语料少之又少，相对而言，未对齐的双语对照语料却多许多。因此，从未对齐的双语对照语料中进行 TP 抽取，是一个有趣而现实的课题。

本文的研究兴趣也集中于后者。

二、抽取算法

Pascale Fung 提出了两种从未对齐双语语料库中抽取翻译词对的方法。第一种方法称为 K 向量(K-vec)法[1]。其思路相当简单，即将每个单语语料库按长度等分成 K 段，单词 w_i 在第 i 段($i=1,2,\cdots,K$)是否出现分别记为 1 和 0，于是每个单词都可以用 K 维布尔向量来表示。通过计算源语言单词和目标语言单词的向量相似性，便可以抽取出相似度很大的源语言和目标语言词对。该方法可以快速、粗略地抽取翻译词对。但是，它的明显缺点就是假定源文本与目标文本长度之间存在着很强的线性关系，于是如果将文本等分成 K 段，可能会对诸如汉英这样的长度线性关系并非十分显著的语言对不合适，况且文本中的任何插入、删除都会进一步加剧文本长度之间的非线性度，从而影响结果的准确性。

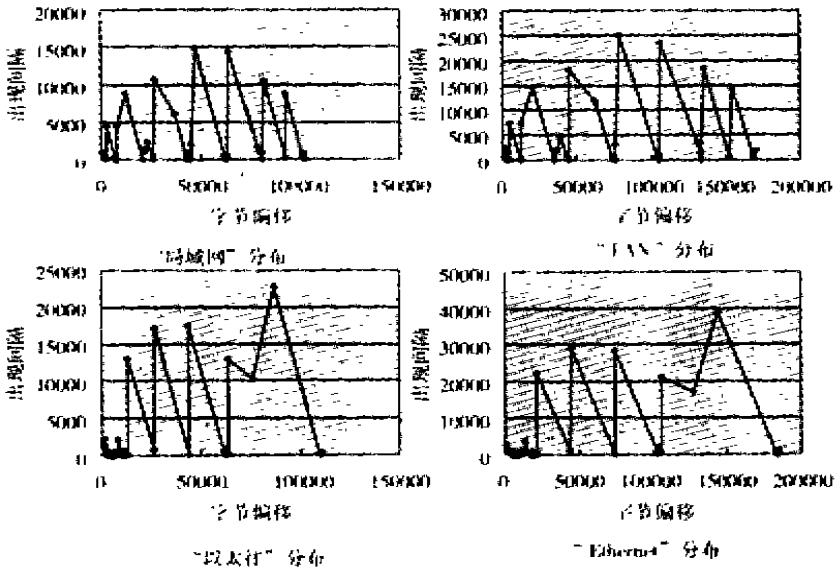

图 1　词分布间隔向量相似性示意图

考虑到上述方法的缺陷，Pascale Fung 又提出了一种基于词对特征匹配(Word Pair Fea-

ture)的称为 DK-vec 的方法[2]。通过考察,Pascale Fung 提取了一种称为"词出现间隔向量"的特征。具体说就是,假定某个词 w 在文本中出现 $n(n>1)$次,它的每次出现所在的字节偏移为 $Offset_1, Offset_2, \cdots, Offset_n$,则它的分布间隔向量为$\langle Offset_2\text{-}Offset_1, Offset_3\text{-}Offset_2, \cdots, Offset_n\text{-}Offset_{n-1}\rangle$。统计发现,互为翻译的高频词的出现间隔向量常常表现出很强的相似性,而不互为翻译的词对的出现间隔向量却常常相似性不强。图 1 中显示的就是从一个汉英对照语料中得到的统计结果。横坐标是词在文本的字节偏移,纵坐标是词的出现间隔。如图可见,"以太网"与"Ethernet"、"局域网"与"LAN"的出现间隔分布十分地相似,而它们之间却差异很大。

于是,通过比较源语言和目标语言词对的出现间隔向量的相似性,同样可以达到抽取翻译词对的目的。由于任意两个词的分布间隔向量的维数可能不同,Pascale Fung 采用了一种称为 DTW(Dynamic Time Wrap)的动态规划算法来实现两个不同维数向量之间的比较。

形式地,假设 $m(m\geq 1)$维向量 X 要和 $n(n\geq 1)$维向量 Y 进行比较,则采用 DTW 算法实现时的递归式为:

$$\begin{aligned} D(X[i:m], Y[j:n]) = \min\{ & d(X[i], Y[j]) + D(X[i+1:m], Y[j+1:n]), \\ & d(X[i], Y[j+1]) + D(X[i+1:m], Y[j+2:n]), \\ & d(X[i+1], Y[j]) + D(X[i+2:m], Y[j+1:n])\} \end{aligned} \quad (1)$$

其中,$X[i:m]$表示由分量 $X[i]$、$X[i+1]$、$\cdots$、$X[m]$组成的向量,$Y[j:n]$ 表示由分量 $Y[j]$、$Y[j+1]$、$\cdots$、$Y[n]$组成的向量,$D(X, Y)$表示 X、Y 两个向量之间的距离(相似度),$d(a,b)$表示向量分量之间的距离。公式(1)中假设匹配时不能漏过多于一个的分量。也就是说,假如当前匹配分量为 $X[i]$与 $Y[j]$,则下一个匹配只可能是 $X[i+1]$与 $Y[j+1]$、$X[i+2]$与 $Y[j+1]$或者 $X[i+1]$与 $Y[j+2]$其中之一。在 Pascale Fung 的方法中,

$$d(a,b) = |a-b| \quad (2)$$

很显然,Pascale Fung 的第二种方法克服了第一种方法的缺点,只需要通过比较词的位置分布信息便可以抽取到可能的翻译词对。即使文本中存在一定的插入和删除章节也不会对结果造成大的影响。当然,实现该算法还有不少的具体细节,详细可以参见文献[2]。

三、改进

我们使用 Pascale Fung 的 DK-vec 方法对一些汉英对照文本进行了实验,但是却达不到文中提出实验的结果。经过反复思考和实验,我们发现原来问题出在公式(2)中。公式(2)在计算向量之间的距离 D 时,利用各个分量距离 d 进行累加计算,而每个 d 又直接采用绝对差值进行计算,我们认为如此做法至少在以下方面存在问题。

(一)归一化问题。通常,互为翻译的源语言文件和目标语言文件长度是不同的,特别对于印欧-非印欧语言对(如英汉),这种长度上的差别更加明显。长度差异的情况下进行距离的直接差值计算,可以认为是在不同数量级上进行计算,在意义上也不成立。

(二)绝对差值和相对差值问题。我们举一个例子来说明这个问题。假设两个互为翻译的词对 A,B,在各自文中均出现 $t+1$ 次,并且一一对应。它们的出现间隔向量分量分别为为 $A_1, A_2, \cdots, A_t$;$B_1, B_2, \cdots, B_t$。如果某个 A_i 和B_i 正好比较大,也就是说 A 和 B 的第 i 次出现和和第 $i+1$ 次出现间隔很大。这样的情况下,A_i 与 B_i 的差值可能相对就很大,使得在所有 $|A_j - B_j|$的运算中$(j=1,2\cdots,t)$,$|A_i - B_i|$起着重要的作用,而显然各个分量差值计算的地

位应该是平等的。也就是说，不管其他差值是如何的小、如何地吻合，$|A_i - B_i|$的巨大差值使得整个距离的计算结果偏大。

（三）距离累加问题。Pascale Fung 在文中计算向量距离时采用的是求累加值的办法。仍以（二）中的例子为例。假设我们只计算 A 和 B 的前 s 项分量的距离 $D_s(A,B) = |A_1 - B_1| + |A_2 - B_2| + \cdots + |A_s - B_s|$，$s < t$。则显然有 $D_s(A,B) < D_t(A,B)$。也就是说，计算的分量距离越多，得到的出现向量距离就越大。按照这种计算，在双语对照文本中，如果两个翻译词对出现的次数越多，计算出的出现向量距离也就越大，这显然是不符合实际要求的。

针对以上问题，我们在算法实现时进行了改进。针对问题（一），我们在计算时引进了一个归一化因子 c。它的值为源语言文本和目标语言文本长度的比值。之所以选用这个长度比值作为归一化因子，不仅仅出于一般归一化问题的考虑，而且也考察了双语文本长度关系的规律性。利用这个规律性，人们还提出了基于长度的双语库句子对齐的方法[3]。针对问题（二），我们采用了相对值的计算来得到分量之间的距离。对于问题（三），我们采用平均值作为最后的度量值。形式地，公式（2）变为：

$$d'(a,b) = |(c * a - b)/a| = |c - b/a| \tag{3}$$

c 为 Y 所在文本长度与 X 所在文本长度的比值。

同时，公式（1）中的递归计算式变为：

$$\begin{aligned} D'(X[i:m], Y[j:n]) = \min\{ & d'(X[i], Y[j]) + D'(X[i+1:m], Y[j+1:n]), \\ & d'(X[i], Y[j+1]) + D'(X[i+1:m], Y[j+2:n]), \\ & d'(X[i+1], Y[j]) + D'(X[i+2:m], Y[j+1:n]) \} \end{aligned} \tag{4}$$

最后的出现向量的距离为：

$$D(X[1:m], Y[1:n]) = D'(X[1:m], Y[1:n])/L_{match} \tag{5}$$

其中，L_{match} 称为向量匹配长度，即计算 D' 过程中真正使用的分量差值个数。计算的过程中同时可以记录匹配路径，匹配路径可以用匹配过程通过的结点表示，而匹配路径的每个结点是指当前选用的匹配分量对，可以用各自的分量号二元组 $\langle x, y \rangle$ 来表示，即当前选用的是 $X[x]$ 与 $Y[y]$ 进行匹配。

与式（4）对应的路径为：

$$\begin{aligned} Path_{i,j} = \langle x_i, y_j \rangle = \operatorname{argmin}_{a,b} (& d'(X[i+a], Y[j+b]) + \\ & D'(X[i+a+1:m], Y[j+b+1:n])) \end{aligned} \tag{6}$$

与之相邻的下一条路径结点为 $Path_{i+x_1, j+y_2}$，计算 $X[1:m]$ 与 $Y[1:n]$ 距离得到的路径的总个数即为 L_{match}。

四、实验结果

我们的实验语料是《计算机世界报》光盘上汉英时文对照选读的文章，总共有 73 篇。我们将汉语文章顺序地放在一个汉语文本中，同时将对应的英语文章放入一个英文文本中。由于原文中拼写错误较多，为了使实验结果不受此因素影响，我们对原文进行了拼写校对。最后得到的汉语文本有 6 万多汉字，英语文本有 3 万多单词，长度各为 135K 和 183K。实验中不仅使用了汉语词，而且还使用了高频汉字接续对，以及高频汉、英语词对。实验中采用了北大计算语言所提供的汉语切分标注工具。汉语词只选用了名词和高频接续对进行匹配。下表列出的是两种方法得到的前 19 项 TP。从表中可以看出，本文提出的方法能够很好地抽取双语对

照文本中的 TP。其准确率也比原来的方法得到很大的提高。

需要指出的是，由于本算法利用的是 TP 的位置相似信息，而显然，位置相似的却不一定是 TP，所以本算法还需要一个后校验过程。举个例子，比如一篇讲"计算语言学"(Computational Linguistics)的文章，就很有可能把"计算/Linguistics"或者"语言学/Computational"也看成 TP。因为它们的出现位置确实也极其相似。其他在单语文本中出现类似的词对在进行匹配时也常常会发生同上面一样的结果，所以对上述算法产生的结果还需要人工或者人机互助的方法加以校验。

另外，表 1 中还有一个部分匹配的问题，即一个片段的一部分与其翻译匹配。这个问题可以转化为字符串之间是否互相包含的判定问题加以解决。这里不再详述。

表 1　两种方法的不同实验结果

序号	原来的方法		改进后的方法	
	汉语单词	英语单词	汉语单词	英语单词
1	文档	document	工具	tool
2	－以太	Ethernet	－以太	Ethernet
3	以太网	Ethernet	以太网	Ethernet
4	－太网	Ethernet	－太网	Ethernet
5	－＊因特	gigabit	协议	protocol
6	＊因特网	gigabit	服务器	server
7	服务器	server	文档	document
8	＊密钥	storage	用户	user
9	－千兆	gigabit	对象	object
10	千兆位	gigabit	客户机	client
11	＊组件	storage	标准	standard
12	对象	object	网络	network
13	用户	user	微软	Microsoft
14	－兆位	gigabit	计算机	computer
15	＊－特网	gigabit	数据	data
16	＊钥匙	storage	交换机	switch
17	数据	data	环境	environment
18	网络	network	－千兆	gigabit
19	系统	system	千兆位	gigabit

＊表示错误匹配　－表示部分匹配

五、结论

本文的算法可以从双语对照文本中抽取出部分 TP，而这些对照文本并不需要进行初始对齐。由于双语对照文本获取相对容易，因此本文算法具有十分广泛而实际的用途。并且，由于算法本身独立于语言信息，所以它同样可以应用于其他语言对之间的 TP 抽取工作。从对照文本中抽取 TP 以后，可以为双语库下一步的加工工作提供基础：第一，在抽取 TP 的同时产生的匹配路径可以作为双语库对齐的初始锚点；第二，抽取的 TP 可以为后续工作提供宝贵的词典资源。

在本文工作的基础上，可以通过加入其他语言信息(如短语构成信息)进行更精确有效的短语 TP 抽取或者术语翻译抽取等等工作。

(下转 57 页)

参 考 文 献

[1] Sergei Nirenburg(editor). The PANGLOSS Mark Ⅲ Machine Translation System A Joint Technology Report, CMU－CMT－95－145, Apr. 1995

[2] 刘志杰. 英汉机器翻译软件长句分析刍议. 见:1999 年计算语言学全国联合学术会议论文集. 北京:清华大学出版社,1999

[3] 王虹. MT 系统的从句分析、设计与实现[硕士学位论文]. 哈尔滨:哈尔滨工业大学计算机系. 1992

[4] 郑杰. 英汉机译系统中的难点分析. 见:1999 年计算语言学全国联合学术会议论文集. 北京:清华大学出版社,1999

[5] 孟遥,赵铁军等. A Decision Tree Based Corpus Approach to English Base Noun Phrase Identificatio. In: Proceedings of the 1st International Conference on East-Asian Language Processing and Internet Information Technology, 沈阳,2000,5－10

[6] 荀恩东. 统计与学习并举的渐进式英语句法分析[博士学位论文]. 哈尔滨:哈尔滨工业大学计算机系,1999

[7] Clausdia Leacock, Geoffrey Towell, Ellen Voorhees. Towards Building Contextual Representations of Word Senses Using Statistical Models. Branimir Boguraev, James Pustejovsky eds. Acquisition of Lexical Knowledge from Text. In: Proceedings of a Workshop sponsored by the ACL, Ohio State University, 1993,10－21

(上接 44 页)

显然,本文得到的工具可以为诸如机器翻译、跨语言检索或者词典编纂等方面的自然语言应用服务。

参 考 文 献

[1] Fung P, Church K W. K-vec: A New Approach for Aligning Parallel Texts. In: Proceedings of the 15th International Conference on Computational Linguistics(COLING'94), Tokyo, Japan, 1994, 1096－1102

[2] Fung P, McKeown K. Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas(AMTA'94), Columbia, MD. 1994, 81－88

[3] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics(ACL'91), Berkeley, CA, 1991, 177－184

[4] Gale W A, Church K W. Identifying Word Correspondences in Parallel Texts. In: Proceedings of the Fourth DARPA Speech and Natural Language Workshop, Pacific Grove, CA, 1991, 152－157

[5] Tiedemann J. Extraction of Translation Equivalents from Parallel Corpora. 1998, From http://stp.ling.uu.se.