第 16 卷第 1 期 电 脑 与 信 息 技 术 Vol.16 No.1
2008 年 2 月 Computer and Information Technology Feb. 2008
文章编号:1005-1228(2008)01-0005-03

新闻领域双语语料建设与句子对齐方法的研究

林哲辉,贾剑锋,郭 文
(厦门大学信息科学与技术学院,福建 厦门 361005)

摘 要:双语对齐的平行语料库在机器翻译和自然语言处理领域中扮演着非常重要的角色,它的研究和制作具有重要的理论意义和实用价值,双语语料的建设十分必要,其中双语对齐是最基本的环节。文章首先简要介绍了语料库的建设情况,然后结合主流的句子对齐方法提出并实现了基于词典和语言学信息的英汉双语句子对齐。

关键词:双语语料库;平行;对齐;机器翻译
中图分类号:TP391 **文献标识码**:A

Building Chinese-English bilingual Corpus of News Field and Research on Sentence Alignment

LIN Zhe-hui, JIA Jian-feng,GUO Wen
(College of Information Science and Technology, Xiamen University,Xiamen,Fujian 361005,China)

Abstract: Bilingual corpus plays a very important part in machine translation and natural language process field, Building and research on it both have theoretic significance and practical value. Therefore, it's necessary to build a large-scale bilingual corpus. Bilingual Alignment is the basic connection. This paper firstly gives a brief discussion on the construction of the corpus, secondly discuss the main ways of sentence alignment, and finally pose another aligning method base on lexical and language information.

Key words: bilingual corpus; parallel; alignment; machine translation

0 引言

国内外很多研究机构都致力于双语语料库的建设,并利用这些语料库进行广泛的研究。但目前汉外双语语料库规模比较小,加工规范也不统一,从而影响了双语语料库知识获取的研究。实现各个层次的对齐是双语语料库建设的一项重要内容。

1 双语语料库建设

加拿大的议会会议录(Canadian Hansards)是非常著名的英法双语语料库,许多最初的基于双语语料库的研究都是在该语料库基础上进行的[1,2]。香港科技大学收集和加工了香港立法委员会的会议记录,形成了汉英双语语料库[3]。目前,我们所拥有的有价值的语料大致如下:

联合国语料库是联合国近几年来会议记录的筛选和整理,包含 1993~2002 年的所有语料。

香港新闻语料也是重要的资源,包含三个子语料库:香港议会平行语料、香港法律平行语料、香港新闻平行语料,分别有 2000 年和 2004 年两个版本,其中 2004 年的版本已经做到句子层级的对齐,但文本仍然混乱,当中也存在不少噪声。所以语料建设的主要工作是句对的抽取和根据句对的评价权重进行筛选。

FBIS(Foreign Broadcasting Information Service)是国外广播信息的语料,包含多国语言的篇章级对齐文本,我们选取其中的中英文本来进行语料库的建设。

汉英新闻杂志平行文本(Chinese English News Magazine Parallel Text)包含的是新闻事件以及它的英文译文,是 LDC 从台湾的 Sinorama 杂志收集的,时间跨度为 1976~2004 年,当中共有 6366 个故事对,365 568 句子对。由于题材和翻译的原因,这部分的语料质量不够好。

2 编码格式规范

收稿日期:2007-11-19;修订日期:2007-12-07
基本项目:福建省自然科学基金项目(2006J0043)
作者简介:林哲辉(1983-),男,福建安溪人,硕士研究生,研究方向为自然语言处理;贾剑锋(1984-),男,新疆伊宁人,硕士研究生,研究方向为机器翻译。

在取得句子级对齐的语料之后，需要用 XML 语言对其进行标注。语料的属性（如篇章、标题、领域、段落编号、句子编号等）在 XML 分别使用不同的标签来描述。因此，我们设计了以下基本标注属性：

<corpus> 表示整个语料库文本，有一些属性分别表示语料库面向的领域（如新闻/法律/议会等）、语料库的语种等。<text> 表示一个文本，有一些属性分别表示文本的原始文件名、原始文件日期等。属性可以动态增加。基本标注属性及说明见表 1（不区分大小写）。

表 1　基本标注属性

XML 标记	属性	意义
Corpus		
	field	领域
	desc	语料库的描述
	textno	语料库中的文本数
	lang	文本语言，由‘-’分割不同语言
Text		文本
	doc	原始文件名
	docid	原始文件编号
	date	原始文件日期
P		段落
	sno	段落中的句子数
	Title	表示是否是标题
S		句子，也称之为 segment，是翻译单位
	Id	句子的顺序号，整个语料库统一编号
Mu		匹配单位
	Lang	语言。当顺序与整个语料库一致时可省略

3　双语语料库句子对齐方法的研究

句子对齐方法可以分为三类：基于长度的方法、基于词汇的方法和混合方法。

基于长度的方法最初由 Brown[1]和 Gale[2]提出，依据是两种语言译文的长度满足一定比例关系，在英法双语的加拿大议会会议录上取得了较好的对齐效果。Chen[4]、Kay[6]分别根据词汇翻译模型和双语单词的分布信息进行了英法双语句子对齐。文献[8]直接利用双语词典对大学英语教材做了句子对齐，也取得了令人满意的效果。基于长度的对齐方法模型简单，独立于语言知识和其他外部资源，但鲁棒性不好，容易造成错误蔓延。基于词汇的对齐方法相对可靠精确，但计算相当复杂。研究人员试图将这两种方法结合起来进行句子对齐。香港大学 Wu[3]通过创建特殊词表来对基于长度的方法进行了改进，并对香港立法委员会的会议记录做了对齐试验，取得了较好的结果。

3.1　形式化描述与概念

句子对齐就是将两种语言中互为译文的句子作为句对放在一起，它的形式化描述为：

设 S 和 T 是互为译文的文本，S 为含有 n 个句子的源语言文本即 $S=s_1, s_2 \cdots s_n$，T 为含有 n 个句子的目标语言文本即 $T= t_1, t_2, \cdots t_m$（s_i 和 t_j 都以句子结束符作为结尾，$i=1\sim n, j=1\sim m$），根据文献[2]的定义，S 和 T 的对齐关系 A(S,T) 可以表示为句珠序列 B：$B=b_1, b_2 \cdots b_k$。

句子对齐的任务就是要找到满足下列条件的一组句珠序列：

（1）任何一个句子属于且仅属于唯一的一个句珠；

（2）任何一个句珠中不再包含更小的句珠；

（3）句子的位置先后和所属句珠的次序是一致的，不允许交叉对应。

设 h 为每个对齐单位（即句珠）的评价函数，则对齐问题可以定义为在对齐的句珠中找到一个最佳对齐序列，使得该序列具有最优的 H 值。形式化描述这样一个优化问题：

$$H(B)=H_h(h(b_1), h(b_2), \cdots h(b_k)) \quad (1)$$

从概率角度出发，句子对齐问题可以看成是要求每种对齐方案中对齐概率最大的 A（S,T），

$$A(S,T) = \arg\max Prob(B) = \arg\max \prod_{i=1}^{k} Prob(bi) \quad (2)$$

句子对齐研究的核心实质就是要找到一个理想的方法来计算句珠的对应程度。

3.2　词形还原

词形还原可以描述为输入英语单词 e，将它还原成原型。我们从互联网获得一个词性转换表，提供了名词、形容词、副词、动词的不规则变换的词和它的原型。我们利用这个词型变换表来查找 e 的原型，如果查不到，则利用 porter stemmer 的规则还原法将 e 还原。

词形还原算法可简单描述如下：

(1) 输入一个单词；

(2) 如果还原词典里有该词，则输出该词转(4)；否则转(3)；

(3) 如果有该词的还原规则，则根据规则调整词形，输出还原后的词；

(4) 如果还有单词则转(1)；否则结束。

3.3　分词与标注

中文分词选用的是厦门大学史晓东老师的 segtag 工具，英文采用的是 Brown 的标注器，使用的是 WSJ 的标注集。

3.4　评价函数的选取

评价函数句对的评价函数的选取是句子对齐方法的关键。它既要与英语句子有关，又要与汉语句子有关,而且还要与二个句子共同有关。我们将英语单词在在词典里的所有译文用于匹配汉语句子，只要英语单词有一个译文和汉语句子中的某个词匹配，则认为这个英语词的译文在汉语句子中存在,反之亦然。另外,我们加入了词性信息的匹配比重。评价 S 和 T 之间关系的评价函数如下:

$$H=\frac{match(Se,Tc)+match(Tc,Se)}{length(S)+length(T)}*\frac{Num(Sn)+Num(Sv)}{Num(Tn)+Num(Tv)} \quad (3)$$

其中,length(S) 和 length(T) 分别代表句对中英语句子和汉语句子的长度,即单词数;match(Se,Tc)代表译文出现在汉语句子中的英语单词数 (根据英汉词典); match(Tc,Se)代表成为英文单词译文的汉语单词数(根据汉英词典); Num(Sn)与 Num(Sv)代表原文中的名词个数和动词个数; Num(Tn)与 Num(Tv)代表译文中的名词个数和动词个数。

3.5　动态规划算法

实验所采取的段落边界都是清晰的，不会影响实验精度。实验前,我们提供了以逗号进行断句的句子序列文本。在文本分词和还原的预处理过程中产生的错误暂不列入考虑范围。在实验中,我们采取了句子对齐研究中经典的动态规划算法来搜索最优路径，主要考虑了 7 种对齐模型,分别为(1- 0) ,(0- 1) ,(1- 1) ,(1- 2) ,(2- 1) ,(1- 3) ,(3- 1) 。

4　实验结果

实验所采取的语料库为 FBIS 的部分段落对齐语料,其中原始文本中的部分段落边界不清晰,实验前已做段落对齐和人工校对的预处理，所以实验中并不包含段落边界错误带来的影响。

实验所采取的规模为 50 个文本段落,其中中文文本包含 864 个句子,英文文本包含 881 个句子。实验结果如图 1 所示:

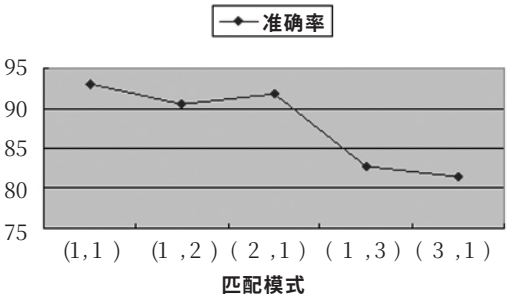

图 1　实验结果

可以看出(1, 1) 匹配模式的句子最常见,对齐准确率最高,其他匹配模式也达到了不错的效果。

5　结论

英汉互译的句对之间的词性信息有着必然的紧密联系,本文从它们之间的这种内在联系出发,提出了基于词典和语言学信息的对齐方法。实验结果证明,这种对齐方法对进行大规模语料建设和整理具有一定的实用意义。

参考文献:

[1] Brown P .F .Aligning sentences in parallel corpora [C].Proceedings of 29th Annual Meeting of the ACL Berkeley, 1991.169- 176.

[2] W. A. Gale, K. W. Church. A Program for aligning sentences in bilingual corpora [J].Computational Linguistics,1993,19(1):75- 102.

[3] Dekai Wu .Aligning a parallel English & Chinese corpus statically with lexical criteria [C]. Proceedings of ACL- 94,1994.82- 87.

[4] M. Kay, M. Rocheisen. Text- translation alignment [J] .Computational Linguistics,1993,19(l):121- 142.

[5] S.F.Chen. Aligning sentences in bilingual corpora using lexical information[C].Proc. of the 31th Annual Meeting of the ACL, 1993.9- 16.

[6] P. Fung, K.W.Church. K- sec, a New Approach for Aligning Parallel Texts[C].Proc. of COLING'94,1994.1096- 1102.

[7] 常宝宝,詹卫东,柏晓静,吴云芳,等.服务于汉英机器翻译的双语语料库和短语库建设[C].第二届中日自然语言处理专家研讨会论文集,2002:147- 154.

[8] 刘昕,周明,黄昌宁.基于长度算法的中英双语文本对齐的试验[A].陈力为.计算语言学进展与应用[C]. 北京:清华大学出版社,1995,62- 68.

[9] 刘昕,周明,朱胜火,黄昌宁.基于自动抽取词汇信息的双语句子对齐[J].计算机学报, 1998, 21(8): 151- 158.