·人工智能及识别技术· 文章编号: 1000-3428(2015)04-0166-05 文献标识码: A 中图分类号: TP391

基于锚点句对的汉维句子对齐方法

塞麦提·麦麦提敏[1],侯 敏[2],吐尔根·伊布拉音[1]

(1.新疆大学信息科学与工程学院,乌鲁木齐 830046;
2.中国传媒大学国家语言资源监测与研究有声媒体中心,北京 100024)

摘 要: 为提高汉维句子对齐方法的准确率,提出一种分段句子对齐方法。采用词汇信息和长度信息相结合的策略,识别出能作为锚点的一对句子(锚点句对),并将其作为分割标志对全文进行分段,在各片段内使用基于长度的方法实现全部句子的对齐,采用词汇、数字、标点符号和长度信息提高方法的领域移植性,使用分段方法避免复杂的计算过程,从而解决错误蔓延问题。实验结果表明,该方法的准确率达到95.2%,比基于长度的句子对齐方法提高了2.7%。

关键词: 平行语料库;句子对齐;锚点;基于长度的方法;基于词汇的方法

中文引用格式: 塞麦提·麦麦提敏,侯 敏,吐尔根·伊布拉音.基于锚点句对的汉维句子对齐方法[J].计算机工程,2015,41(4):166-170.

英文引用格式: Saimaiti Maimaitimin,Hou Min,Tuergen Yibulayin. Chinese-Uyghur Sentence Alignment Method Based on Anchor Sentence Pairs[J]. Computer Engineering,2015,41(4):166-170.

Chinese-Uyghur Sentence Alignment Method Based on Anchor Sentence Pairs

Saimaiti Maimaitimin[1],HOU Min[2],Tuergen Yibulayin[1]

(1. College of Information Science and Engineering,Xinjiang University,Urumqi 830046,China; 2. National Broadcast Media Language Resources Monitoring & Research Center,Communication University of China,Beijing 100024,China)

【Abstract】 The step-by-step sentence alignment method is introduced in order to improve current Chinese-Uyghur sentence alignment method. Lexical and length information is used to generate some anchor sentences. Texts are divided into several sections by using anchor sentence as boundary,and then sentences in each section are aligned using length-based method. This method is effective in multi domain text because it uses words,numbers,and punctuation marks. It avoids complex computing and error spreading because of its "subsection" technique. Experimental results show that the precision of this method is 95.2% in Chinese-Uyghur multi-domain texts,which is 2.7% higher than length-based method.

【Key words】 parallel corpora; sentence alignment; anchor; length-based method; lexical-based method

DOI: 10.3969/j.issn.1000-3428.2015.04.031

1 概述

建立具有应用价值的大型双语语料库最重要的技术之一是自动句子对齐。所谓句子对齐是找出原文句子在译文中对应翻译句子的过程或者结果。句子对齐是双语语料库加工的一个重要课题,也是进一步利用平行语料库获取一些双语知识的必要前提。目前,常用的句子对齐方法有基于长度的方法、基于词汇信息的方法和混合的方法。其中,基于长度的方法和基于词汇信息的方法都有缺点,只有混合的方法才能够更好地实现汉维双语句子对齐。然而,混合法也有不同的实现方式,有的以长度信息为主,先进行初步对齐,然后用词汇信息,对初步对齐结果进行进一步处理和筛选,从而提高句子对齐的准确率;有的则以词汇信息为主,长度信息为辅;混合方法在不同应用系统中所参考的词汇信息也不同。

在汉维句子对齐研究方面,文献[1]通过大量训

基金项目:新疆维吾尔自治区自然科学基金资助项目(2012211B08)。

作者简介:塞麦提·麦麦提敏(1980-),男,讲师、博士,主研方向:自然语言处理;侯 敏、吐尔根·伊布拉音,教授、博士生导师。

收稿日期:2014-05-12 修回日期:2014-06-06 E-mail:tilchin@126.com

练语料,对汉维平行文本间存在的长度规律进行研究,提出了基于长度的汉维句子对齐方法。文献[2]采用基于锚点的方法进行汉维双语句子对齐实验。根据其实验结果,基于锚点的句子对齐方法在汉维双语句子的对齐中达到了 85.2% ~89% 的准确率。文献[3]利用回车符和数字信息相结合的多层次分段对齐方法,实现了基于词典翻译的限定领域双语句子对齐。文献[4]采取多策略的方法改进了汉维句子对齐方法。但是,他们的研究主要集中在政府文献等限定领域文本中的句子对齐,准确率也不够高,因此,找到一个高效可行的混合方法已成为汉维语句子对齐研究的重要任务。

本文提出一种基于锚点句对(可以作锚点的一对句子)的句子对齐方法,并在汉维平行语料库建设中使用。该方法将句子对齐问题转化为"分段 + 对齐"的问题,在"分段"过程中利用专有名词、术语、标点符号、数字等信息,抽取一批锚点句对,并以此作为分割标志将全文分成互相对齐的片段,然后在各片段内采用基于长度的方法进行句子对齐。其中,锚点句对的抽取是该方法的关键,也是研究的重点。

2　锚点句对及其特点

双语句子对齐除了 1-1 对齐模式以外,还有1-2,2-1,2-2 等多种对齐模式,其中,1-1 模式占绝对多数。从理论上讲,任何模式的句对都可成为锚点句对,但是在实际操作中考虑到方法的简易性,只将 1-1模式的句对作为候选锚点。因此,本文方法中锚点句对是指分别属于原文和译文,满足一定条件的 1-1对齐模式的句对。本文引入双语文本示意图(图 1),以便表述锚点句对。

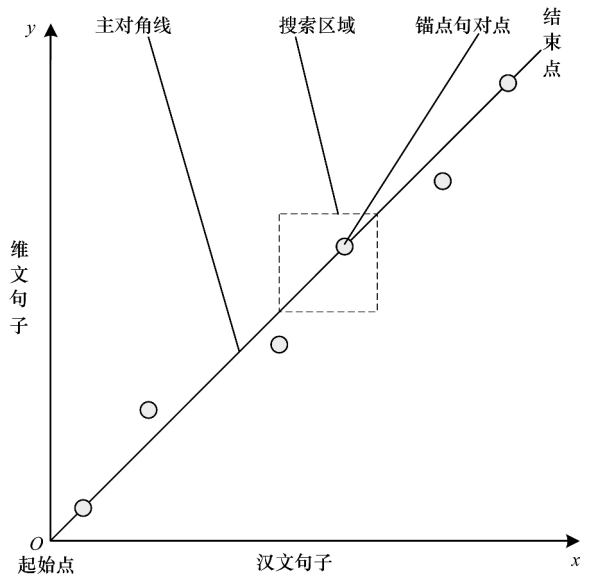

图 1　双语文本示意图

一般的双语文本图中 x 轴、y 轴一般以字节数为单位,分别表示原译文的长度位置,而在本文的双语文本图中,x 轴、y 轴则以句子为单位,它们分别用来表示汉文句子($C_1, C_2, \cdots, C_m$)和维文句子($U_1, U_2, \cdots, U_n$)。

在图 1 中,起始点和结束点间的连线即为双语图的主对角线,主对角线的斜率称为双语图的斜率。用 $Anchor(C_m, U_n)$ 来表示由汉文第 m 个句子和维文第 n 个句子构成的锚点句对。根据双语图,锚点句对具有以下特点:

(1) 线性特征: 所有锚点句对点在双语文本图趋于一条直线。

(2) 斜率特征: 由所有锚点句对点形成的最小平方线一般不会出现与双语文本图的主对角线斜率相差很大的情况。

(3) 单调性: 任意 2 个锚点句对点不会有相同的 x 或 y 坐标值。若已知一个锚点句对 $Anchor(C_m, U_n)$,则一定不会存在另一个锚点句对 $Anchor(C_x, U_y)$,其中 $m = x$ 或 $n = y$。

(4) 递增性: 若已知一个锚点句对 $Anchor(C_m, U_n)$,则一定不会存在另一个锚点句对 $Anchor(C_x, U_y)$,其中,$m > x$ 而 $n < y$ 或者 $m < x$ 而 $n > y$。

若在确定锚点句对过程中出现违反上述锚点句对特性的句对,则称锚点句对存在冲突,就需要通过一定的策略排除这种冲突句对,识别正确的句对。

3　锚点句对抽取方法

如图 2 所示,本文提出的基于锚点句对的句子对齐方法,首先利用词汇、数字、标点符号和长度等信息生成一批锚点句对,并以锚点句对作为分割标志将全文分成若干个更小的互译片段,然后在每个片段内采用基于长度的方法实现汉维双语句子的对齐。

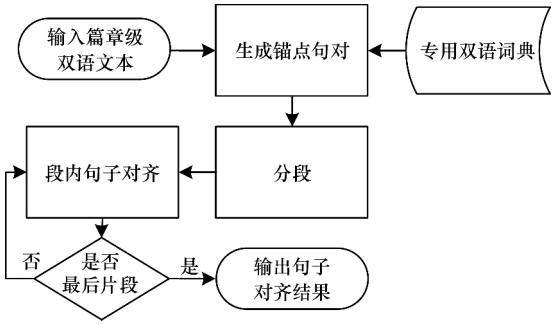

图 2　基于锚点句对的句子对齐方法流程

在图 2 中的段内句子对齐阶段,采用类似于文献[5]基于长度的方法,首先利用简单的统计模型,获得汉维句子对齐的相关参数值,然后利用动态规划的方法完成句子对齐。需要指出的是本文基于长

度的方法分别将汉语字符和维语单词数作为句子长度计算单位。基于锚点句对句子对齐方法的关键是锚点句对抽取方法。锚点句对方法的 2 个重要参数是词汇匹配度和长度相似度。

3.1 词汇匹配度的计算

作为分段的界标，锚点句对必须具有很高的准确率，如只使用长度信息可能会造成很大的误差，解决方法可以是利用覆盖率较高的双语词典，对汉维句子中的所有词汇进行匹配，求全部词汇的互译率。但是，为了计算互译率，需要进行汉文分词，维文词干还原等操作，因此，求全部词汇的互译率是时间开销很大的过程。此外，同一个汉语单词可以有多种维语翻译，这就导致这种方法更加低效。对此，本文采用只求部分词汇互译率的策略。选取出现频率较高、翻译固定的以下词语(符号)作为重要线索：(1)专有名词：人名，地名，机构名；(2)术语：常用的专业术语；(3)数字：阿拉伯数字，汉文大小写数字，维文数字；(4)标点符号：汉语和维语的标点符号；(5)特殊字符：如@ &MYM = #等多语种都相同的符号字符。其中，专有名词和术语具有词义单一性和翻译对应性特点，其在原文和译文中的对应关系非常明显，使用频率也高；数字、标点符号和其他特殊字符也有如表 1 和表 2 所示的对应关系，在双语句子中以相同(相近)的形式出现，计算机对它们的识别比较容易。

表 1 汉语和维语数字对应表

汉文形式	维文形式	汉文例子	维文例子
阿拉伯数字	阿拉伯数字	168	168
阿拉伯数字	单词表达	168	بىر يۈز ئاتمىش سەككىز
大(小)写数字	阿拉伯数字	三十五	35
大(小)写数字	单词表达	三十五	ئوتتۇز بەش
阿拉伯数字和大(小)写数字	阿拉伯数字	5 千(仟)	5 000
阿拉伯数字和大(小)写数字	单词表达	5 千(仟)	بەش مىڭ
阿拉伯数字和大(小)写数字	阿拉伯数字和单词	5 千(仟)	5مىڭ

表 2 汉语和维语常用标点符号对应表

标点符号	汉文	维文	标点符号	汉文	维文
句号	。	.	分号	；	؛
问号	？	؟	冒号	：	:
叹号	！	!	引号	“ ”	“ ”或 "
逗号	，	،	括号	（）	（）
顿号	、	،	书名号	《》	« »

在词汇匹配度的计算方法上，首先建立了分别收录上述 5 类词汇(符号)的句子对齐专用双语词典。该词典目前收录出现频率较高的 6 000 多条词语，这些词语是从各类汉维双语词典和新疆大学“汉维双语平行语料库”中抽取的。然后，在词典的基础上，通过识别匹配过程，对汉维双语句子中的相关词汇(符号)进行匹配度的计算。汉维句对中词语匹配度的计算公式如下：

$$MatchScore(i,j) = 1 - \frac{(x-y)^2}{(x+y)^2} \tag{1}$$

其中，x 是维文句子词语总数；y 是成功匹配的维文词语总数；$x \geqslant 1$，$y \geqslant 0$，并且 $x \geqslant y$。因此，汉维句子 $S(i,j)$ 的词汇匹配度 $MatchScore(i,j)$ 的取值范围为 0～1。如果在句子中的全部词语匹配成功，则 $x = y$，$MatchScor(i,j) = 1$(最大值)，如果没有词语匹配，则 $y = 0$，$MatchScore(i,j) = 0$(最小值)。

3.2 长度相似度的计算

长度计算单位有字符数、词语数等多种。文献[6]按照词数计算句子长度，而文献[7]依据字母数量计算句子长度。文献[1-3]在汉维句子对齐中都用字符数计算汉维句子长度。以字符作为长度单位进行句子对齐是拼音文字之间可以采取的方法，但是维文是拼音文字，汉文则不是。汉文字符和维文字符属于不同层面的语言单位，其功能和特点有较大差距，以字符作为句子长度单位不太适合于汉维这 2 种文字的特点。单词数作为长度单位进行统计也存在着较大的缺点。在汉维句子中，确实不宜同时采用词数(或字符数)作为句子长度的计算单位，而汉语字符和维文词语的功能基本相同，相关统计结果也表明，互译的句子中的汉文字数和维文词数具有高度相关性，句子长度比值更近似于正态分布。

在本文方法中，分别将汉文字数和维文单词数作为句子长度计算单位，进行句子长度的计算。句子长度相似度可用如下计算式获取：

$$LenSimilar(C_i,U_j) = \frac{|Len(C_i)/Len(U_j) - C|}{C} \tag{2}$$

其中，C 为汉维双语文本中每个维语单词所对应的汉语字符数的期望值，从文献[8]获得汉维句子长度比值的数学期望值($C = 2.01$)；C_i 表示汉文句子序列；U_j 表示维文句子序列；$LenSimilar(C_i,U_j)$ 表示汉维句对 (C_i,U_j) 的长度相似度。但是，汉维句子较短时，它们之间的长度关系并不十分稳定，所以，在方法中需加入调节因子 FL，以避免在计算短句间长度关系相似度时造成较大误差，则有：

$$LenSimilar(C_i,U_j) = \frac{|Len(C_i)/Len(U_j) - C|}{C} \times FL \tag{3}$$

其中,如果 $Len(C_i) \leqslant StableLen$,则有 $FL = a \times Len(C_i)/StableLen$,否则,$FL = 1$。其中,$a$,$StableLen$ 均是常数,根据文献[9],汉语句子长度大于 10 个字符时,汉维句子长度关系相对稳定,因此,令 $StableLen = 10$,a 为常数,计算得出 $a = 0.5$。

此外,考虑到某些汉维双语文本间的长度关系与统计得到的经验值 C 相差较大的情况,对上述算式进行改造如下:

$$LenSimilarC_i, U_j) = \frac{|Len(C_i)/Len(U_j) - C'|}{C'} \times FL \quad (4)$$

其中,C'为当前维文文本长度与汉文文本长度比值和统计得到的数学期望值 C 的一个平均值,可用式(5)计算:

$$C' = (C + Len(S)/Len(T))/2 \quad (5)$$

其中,$Len(S)/Len(T)$ 是当前汉文文本和维文文本的长度比值。

3.3　方法描述

锚点句对抽取方法采取多层次的识别筛选策略,充分利用词汇(符号)和长度信息,尽可能地获得高准确率的锚点句对。在本文方法中,确定一个锚点句对,需经过以下 3 个阶段:

(1) 产生阶段

产生阶段从在双语文本图中选择一个很小的矩形区域开始,并且这个矩形区域的对角线与双语图的主对角线平行。在这个矩形搜索区域中,锚点句对识别模块将查找所有满足条件的候选句对,如果没有候选句对,搜索区域将适当扩大,直到在这个搜索区域内找到一组以上的候选句对。

在本文方法中,判断是否为候选句对,需要同时满足以下条件:

1) 长度关系

汉维句子 $S(i,j)$ 的长度应满足:

$$MaxLenRatio > LenRatio(i,j) > MinLenRatio \quad (6)$$

其中,$LenRatio(i,j)$ 是汉维句对 $S(i,j)$ 的长度比值。最大长度比值 $MaxLenRation$ 和最小长度比值 $MinLenRatio$ 的计算公式如下:

$$MaxLenRatio = C' + A/(L_j + B)$$
$$MinLenRatio = C' - A/(L_j + B) \quad (7)$$

其中,A,B 均为常数,根据统计实验取 $A = 10$,$B = 14$;L_j 为汉语句子的字数;C'的取值可用式(5)计算。

2) 最大背离角度

如果句对点 $S(i,j)$ 满足上述条件 1),则进行最大角度背离(Maximum Angle Deviation,MAD)的检查。若 $S(i,j)$ 与前面已识别的锚点句对点形成的最小平方线与双语图主对角线之间的夹角小于最大背离角度阈值(暂定为 4),则 $S(i,j)$ 作为候选句对点,否则 $S(i,j)$ 不能成为候选句对点。其中,设最小平方线斜率为 A,主对角线斜率为 B,可用式(8)求得它们之间的夹角:

$$\theta = \arctan\left(\frac{|A - B|}{1 + A \times B}\right) \quad (8)$$

3) 词汇匹配度

词汇匹配度是衡量双语句子成为候选锚点句对的重要依据。

若 $S(i,j)$ 的词汇匹配度 $MatchScore(i,j) > t$,则判断该句对为候选锚点句对,否则不能成为候选锚点句对。其中,t 为阈值(暂时确定为 0.5),可用式(1)求得 $MatchScore(i,j)$。

产生阶段所生成的锚点句对具有较高的准确率,但是在某种情况下还会产生错误,需加入修正部分,以便减少错误,进一步提高准确率。

(2) 修正阶段

在修正阶段,排除一些存在冲突的或不可能为锚点的句对;主要检查候选句对点是否存在违反锚点句对单调性和递增性特性的句对。若存在问题,则排除该句对作为候选句对的可能性。

(3) 识别阶段

通过前 2 个阶段的多层次筛选,在目前的搜索区内获得一定数量的候选锚点句对之后,识别阶段的作用是词汇匹配度和长度相似度作为参数,从候选句对中选取唯一、最佳锚点句对。判定最佳锚点句对的函数为:

$$S(i,j) = Arg\ Max((MatchScore(C_i,U_j) \times e_1 + LenSimilar(C_i,U_j) \times e_2) \quad (9)$$

其中,词汇(符号)匹配度 $MatchScore(i,j)$ 和长度相似度 $LenSimilar(C_i,U_j)$ 分别可用式(1)、式(4)计算。e_1 和 e_2 是加权值,在文献[8-9]基础上,对 1 300 个句子进行句子对齐实验,对相关参数进行测试获取 $e_1 = 0.667$,$e_2 = 0.333$。

4　实验结果与分析

从“新疆大学汉维双语平行语料库”中,随机抽取分别属于文学、法律、公文、学术、新闻等 5 种语体的 10 个文本(一共 1 482 句对)作为测试语料。分别采用基于长度的和基于锚点句对的对齐方法进行实验比较这 2 种方法在多领域的汉维双语文本的句子对齐准确率。

实验评价标准如下[10-12]:

$$\text{准确率} = \frac{\text{正确识别的句对总数}}{\text{识别的句对总数}} \times 100\%$$

$$\text{召回率} = \frac{\text{正确识别的句对总数}}{\text{语料句对总数}} \times 100\%$$

实验结果如表 3、表 4 所示。

表 3 基于长度的句子对齐实验结果

语体	正确识别总数	识别总数	语料句对总数	准确率/%	召回率/%
文学	292	330	335	88.5	87.2
法律	357	362	364	98.6	98.1
公文	244	250	252	97.6	96.8
学术	336	382	378	88.0	88.9
新闻	134	150	153	89.3	87.6
合计	1 363	1 474	1 482	92.5	92.0

表 4 基于锚点句对的句子对齐实验结果

语体	正确识别总数	识别总数	语料句对总数	准确率/%	召回率/%
文学	310	335	335	92.5	92.5
法律	364	364	364	100.0	100.0
公文	248	250	252	99.2	98.4
学术	345	382	378	90.3	91.3
新闻	143	150	153	95.3	93.5
合计	1 410	1 481	1 482	95.2	95.1

根据实验结果，基于锚点句对的句子对齐方法的准确率和召回率分别为 95.2% 和 95.1%，与基于长度的句子对齐方法相比，准确率提高了 2.7%。

通过分析实验结果，可得出以下结论：

(1) 基于锚点句对的句子对齐方法适合于汉维语这样差异较大的语言，其在多领域文本中的准确率也较高。

(2) 在不同领域文本中，句子对齐的准确率也不相同。法律文献等忠实原文、对应性较高的文本中，准确率较高，甚者可以达到 100%。在新闻文本中，基于锚点句子的方法，相对于长度的方法而言，准确率更高。

(3) 文本中译文缺失、翻译模式复杂、句子切分不准等情况会影响句子对齐的准确率。通过错误对齐的分析发现，本文实验的新闻文本中，两处的译文缺失，影响了准确率。此外，文学文本中存在 1-3 模式的句对。由于目前的方法没有考虑 3-1 等类型的句对，无法进行正确对齐。

(4) 基于锚点句对的方法有效提高了句子对齐的准确率，但是一旦某一锚点句对的识别出现错误，将直接影响后续句子的正确对齐。因此，采取有效措施，保证锚点句对的准确率是该方法的关键。

5 结束语

本文提出一种基于锚点句对的句子对齐方法。生成锚点句对，以锚点句对作为分割标志，对全文二次划分并进行句子对齐。实验结果表明，该方法不像基于词汇的方法需要完整的双语词典，其运行效率较高，由于采取了分段对齐策略，可解决基于长度方法的错误蔓延问题，利用长度信息和部分词汇(符号)信息，可用于多种载体文本的句子对齐。

参考文献

[1] 毕雪华. 汉维双语语料库中句子对齐技术的研究[D]. 乌鲁木齐: 新疆大学，2006.

[2] 牛洪梅. 服务于汉维机器翻译系统的双语句子对齐的研究[D]. 乌鲁木齐: 新疆大学，2007.

[3] 热西旦. 汉文-维吾尔文双语语料库构建的实验性研究[D]. 乌鲁木齐: 新疆大学，2007.

[4] 田生伟，吐尔根·依布拉音. 多策略的汉维句子对齐[J]. 计算机科学，2010，37(4)：215-218.

[5] Gale W，Church K. A Program for Aligning Sentences in Bilingual Corpora[C]//Proceedings of the 29th Annual Meeting of ACL. Stroudsburg，USA：Association for Computational Linguistics，1991：177-184.

[6] Brown P F，Mercer R L. Aligning Sentences in Parallel Corpora[C]//Proceedings of the 29th Annual Meeting of ACL. Stroudsburg，USA：Association for Computational Linguistics，1991：169-176.

[7] Gale W，Church K. A Program for Aligning Sentences in Bilingual Corpora[J]. Computational Linguistics，1993，19(1)：75-90.

[8] Mamitimin S. Chinese-Uyghur Sentence Alignment：An Approach Based on Anchor Sentences[C]//Proceedings of the 2nd Workshop on Building and Using Comparable Corpora：From Parallel to Non-parallel Corpora. Singapore：Association for Computational Linguistics，2009：38-45.

[9] 塞麦提·麦麦提敏. 汉维平行语料库构建研究[D]. 北京: 中国传媒大学，2009.

[10] 李维刚，刘 挺，张 宇，等. 基于长度和位置信息的双语句子对齐方法[J]. 哈尔滨工业大学学报，2006，38(5)：689-692.

[11] 祝志杰. IHSMTS 中汉英双语句子对齐机制的设计与实现[D]. 南京: 南京理工大学，2002.

[12] 张 艳，柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报，2005，19(5)：31-37.

编辑 刘 冰