古汉语与现代汉语句子对齐研究

刘 颖 王 楠
(清华大学中文系 北京 100084)

摘 要 对西汉时期司马迁所著《史记》原文(古文)与现代文译文(现代文)的平行语料进行句子对齐研究。对数线性模型将句子的长度特征、句子对齐模式特征和共现汉字特征相结合来对《史记》古文和现代文进行句子对齐。通过实验可以看出,同时考虑句子长度、句子对齐模式和共现汉字三个特征,句子对齐的准确率和召回率是最高的,准确率为 94.4%,召回率为 94.3%。

关键词 句子对齐 对数线性模型 对齐模式

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2013.11.036

RESEARCH ON CLASSICAL AND MODERN CHINESE SENTENCE ALIGNMENT

Liu Ying Wang Nan
(*Department of Chinese Language and Literature, Tsinghua University, Beijing 100084, China*)

Abstract Sentences alignment for Chinese parallel corpus is studied in the paper. The parallel corpora are the original text (classical Chinese) and its modern text translation (modern text) of Shiji (Records of the Grand Historian) written by SiMa Qian in the period of Western Han Dynasty. The log-linear model combines the length feature and sentence alignment mode feature of the sentence with the co-occurrence of Chinese words feature, in this way to align the sentences of the classical Chinese and the modern text of Shiji. Through the experiment it can be demonstrate that the precision and recall rate of sentence alignment reach the highest at 94.4% and 94.3% respectively when taking into account these three features at the same time.

Keywords Sentence alignment Log-linear model Alignment mode

0 引 言

双语对齐是机器翻译研究的重要组成部分。所谓“双语对齐”,就是在翻译过程中源语言和目标语言在语言单位上的各种对应关系。根据语言单位的不同,可以将双语对齐按照语言单位的大小分为:段落对齐、句子对齐、组块对齐、短语对齐和词汇对齐[11]。

从 20 世纪 80 年代开始 Brown[1]、Gale 和 Church[2]、Kay 和 Roscheisen[3]、Smard[4]、Chen[5] 等学者就对双语句子对齐展开了研究。句子对齐的方法主要有基于长度的方法和基于词汇对齐的方法[11]。

基于长度的方法:源语言和目标语言间句子的翻译顺序不存在变化的前提下,两种语言的互译句子在长度上高度相关,即一种语言中的长句翻译成另外一种语言后,仍为长句;一种语言中的短句翻译成另外一种语言后,仍为短句[1,2]。基于长度的句子对齐方法最初由 Brown[1] 和 Gale[2] 提出。基于长度的句子对齐与对齐模式和句子长度相关。Brown 是以单词数为句子的长度单位,利用双语同时存在的特定注释作为锚点,锚点之间的文字片段就能够一一对应,进而缩小文字匹配的篇幅,提高准确率。Gale 以字符数为句子的长度单位。

基于词汇的对齐方法[3-5] 主要有两种:(1)利用双语中的特殊符号和特殊单词进行句子对齐工作,如标点符号、数学符号、人名、地名和机构名等;(2)利用同一语系语言之间存在的同源词或机器词典来进行句子对齐。Kay 和 Roscheisen 利用部分词语的对齐来指导句子对齐,提高了句子对齐的准确率[3]。Simard 利用英语和法语的一些同源词进行句子对齐[4]。

在汉语研究方面,Wu 则利用特殊词汇作为句子长度单位,对香港立法委员会的会议记录进行了实验,取得较好的效果[6]。Wang 根据中文和日文的特点,利用共现汉字特征对基于长度的句子对齐方法进行了改进,在中文和日文句子对齐过程中取得了较好的效果[7]。王斌系统地比较了基于长度和句子对齐、基于词汇的句子对齐和两种结合的方法,并给出结论:基于长度和句子对齐准确率高于基于词汇的句子对齐准确率,并且两种方法结合会比较好[11]。吕学强针对汉英法律语料库进行基于长度的句子对齐,同时对长度评价函数和标准因子作出了符合汉英特点的新定义[12]。林准[13] 和郭锐[14] 则对古文和现代文的句子对齐工作进行了尝试。

对《史记》原文和译文进行了一些分析,发现原文和译文对应的语序基本保持一致;句子对齐与句子长度高度相关,即古文长句与现代文长句对应,古文短句与现代文短句对应;对齐模式多以 1-1 模式为主,其他模式较少;对齐的古文与现代文中存在着较多的相同汉字;人名、地名等专有名词在古文和现代文中

收稿日期:2012-06-06。教育部自主科研项目(20111081010)。
刘颖,副教授,主研领域:自然语言处理。**王楠**,硕士生。

保持不变。正是基于古文—现代文对齐句子的这些特点来设计句子对齐模型特征。也就是,古文与现代文句子对齐与句子长度、对齐模式和共现汉字都有关系,根据古文和现代文翻译的这些特征,采用了对数线性模型。对数线性模型可以考虑多个特征(特征数可以多于两个),而 Gale 等人采用的是隐马尔科夫模型(HMM)只能考虑两个语言特征。隐马尔科夫模型是对数线性模型的特例。当对数线性模型也考虑两个语言特征时,对数线性模型就是隐马尔科夫模型。从古文与现代文句子对齐规律来看,考虑的语言信息越多,句子对齐的可能性越大。因此,对数线性模型更适合作为古文和现代文句子对齐的模型。

与其他双语句子对齐相比,一般双语句子对齐采用的是句子长度特征和对齐模式特征、或词语互译特征和对齐模式特征。句子长度采用词语个数或字符个数。而古文和现代文句子对齐采用三个特征,且句子长度采用共现汉字个数。古文和现代文对齐句子长度平均值和方差值与其他语言对之间的句子长度平均值和方差值不同,同时,句子对齐模式的概率与其他语言对之间也不同,古文和现代文的共现汉字与日语和汉语之间的共现汉字多少和规律也不相同。虽然古文和现代文使用的对数线性模型与其他语言对之间采用的模型有相同的地方,但古文和现代文的三个特征值与其他语言对是不同的,这些值必须根据大规模双语句子对齐语料库来进行统计。

古文和现代文句子对齐与具有同源词或共现汉字的两种语言句子对齐相同,可利用共现汉字进一步帮助判断句子是否对齐。但与未有同源词的两种语言句子对齐不同。

1 对数线性模型

Och 和 Ney 提出的对数线性模型易于整合各种特征[9],在统计机器翻译中得到了广泛应用。刘群和刘洋使用对数线性模型进行词汇对齐,准确率较高[10]。

句子对齐可以定义如下:e、f 表示源语言和目标语言进行篇章或段落对齐后的语料,e 和 f 分别由句子构成,表示成为 $e = e_1e_2\cdots e_I$ 和 $f = f_1f_2\cdots f_J$。如果句子 e_i 和 f_j 互译或 e_i 和 f_j 部分互译,则用 (i,j) 表示这种互译关系。用 a 表示 e 与 f 的一种对齐方式,则句子对齐过程就可看作求 $Pr(a|e,f)$ 最大值的过程。

$$Pr(a\mid e,f) = \frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(a,e,f)\right]}{\sum_{a'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(a',e,f)\right]} \tag{1}$$

在对齐过程中,我们引入 M 个特征函数 $h_m(a,e,f)$,$m = 1,2,\cdots,M$。对于每一个特征函数有一个系数 λ_m,作为特征函数的权重。

则我们所得的对齐结果 $\hat{a}$ 满足:

$$\hat{a} = \operatorname{argmax}_a\sum_{m=1}^{M}\lambda_m h_m(a,e,f) \tag{2}$$

本文采用的模型选用长度特征、对齐模式特征和共现汉字特征 3 个特征。其中长度特征源于 Gale 和 Church[2] 提出的基于长度的对齐模型。对齐模式特征则重点考虑对齐模式的概率,而共现汉字特征则考虑古文句子与现代文句子相同汉字的数量。现代汉语词汇以 2 字词为主,多是由古代 1 字词发展演变而来,在互译的古文词汇和现代文词汇中,有很多现代文词汇包含古文词汇,使得在计算过程中不必再使用额外的词典,而是通过直接计算古文句子与现代文句子相同汉字的数量,以此为特征来判断古文与现代文是否应该对应。

假设句子的一种对齐方式 $a = \{a_1,a_2,\cdots,a_K\}$,$a_K$ 表示 e 与 f 中的第 k 组对齐句子,那么句子对齐模型为如下形式:

$$\begin{aligned}\hat{a} &= \operatorname{argmax}_a\Big(\lambda_1\sum_{k=1}^{K}L_k(a,e,f) + \lambda_2\sum_{k=1}^{K}M_k(a,e,f) + \lambda_3\sum_{k=1}^{K}H_k(a,e,f)\Big)\\ &= \operatorname{argmax}_a\sum_{k=1}^{K}d_k(a,e,f)\end{aligned} \tag{3}$$

式(3)中的 $L_k(a,e,f)$ 为第 k 组对齐句子的长度特征,本文以汉字为单位计算句子的长度。本文长度特征表示如下:

$$L_k(a,e,f) = -100\times\log2\left(1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\delta}e^{-\frac{z^2}{2}}dz\right) \tag{4}$$

$c = \dfrac{\sum l_2}{\sum l_1}$ 表示双语对齐句子的长度平均值,$s^2 = D\left(\dfrac{(l_2 - c\cdot l_1)}{\sqrt{l_1}}\right)$ 表示 $\dfrac{(l_2 - c\cdot l_1)}{\sqrt{l_1}}$ 的方差。则公式(4)中的 $\delta(l_1,l_2) = \dfrac{(l_2 - c\cdot l_1)}{\sqrt{l_1 s^2}}$ 满足标准正态分布。其中,l_2 和 l_1 分别为现代文和古文对应句子的长度。本文实验通过计算得出 $c \approx 1.617$,$s^2 \approx 1.56$。

$M_k(a,e,f)$ 为第 k 组对齐句子的对齐模式特征。我们对人工对齐的部分《史记》语料进行统计,计算出各种对齐模式出现的频率如表 1。根据表 1,句子对齐模式 1-1,1-0,0-1,1-2,2-1 和 2-2 占 98.98%,所以本文主要考虑这 6 种句子对齐模式。对齐模式特征计算如下:

$$M_k(a,e,f) = -100\log(\Pr(m-n)) \tag{5}$$

表 1 句子对齐模式特征统计

	类别(m-n)	频率	概率	
主要类型	0-1,1-0	4	0.15%	98.98%
	1-1	2480	90.12%	
	1-2	106	3.85%	
	2-1	125	4.54%	
	2-2	9	0.33%	
其他类型	1-3	11	0.40%	1.02%
	1-4	4	0.15%	
	2-4	1	0.04%	
	3-1	8	0.29%	
	3-2	3	0.11%	
	5-1	1	0.04%	
总计		2752	100%	100%

$H_k(a,e,f)$ 为第 k 组对齐句子的共现汉字特征。共现汉字特征主要用于衡量古文与现代文中句子之间相同汉字出现的情况。如果两个句子中出现相同汉字所占比例较高,则说明这两个句子对应的可能性很大。根据 Wang[7] 提出的算法计算汉字特征值,见式(6)。

式(7)—式(9)中,$C(s_i)$ 和 $C(t_u)$ 分别表示 s_i 和 t_u 的汉字数,$C(s_i\cap t_u)$ 表示 s_i 和 t_u 相同的汉字数,$C(s_i\cap s_j\cap t_u)$ 表示 s_i、s_j 和 t_u 相同的汉字数,$h_1(i,u;0,0)$、$h_2(i,u;j,0)$ 和 $h_3(i,u;0,v)$ 分别按照式(7)—式(9)进行计算。

$$H(i,u;j,v)=\begin{cases}0 & 0-1\ \text{or}\ 1-0\\ h_1(i,u;0,0) & 1-1\\ h_1(i,u;0,0)+h_2(i,u;j,0) & 2-1\\ h_1(i,u;0,0)+h_3(i,u;0,v) & 1-2\\ \min\begin{cases}h_1(i,u;0,0)+h_2(i,u;j,0)+h_1(0,0;j,v)\\ h_1(i,u;0,0)+h_3(i,u;0,v)+h_1(0,0;j,v)\\ h_1(i,0;0,v)+h_1(0,u;j,0)\end{cases} & 2-1\end{cases}\tag{6}$$

$$h_1(i,u;0,0)=\frac{C(s_i\cap t_u)}{\min(C(s_i),C(t_u))}\tag{7}$$

$$h_2(i,u;j,0)=\frac{(C(s_j\cap t_u)-C(s_i\cap s_j\cap t_u))}{C(s_j)}\tag{8}$$

$$h_3(i,u;0,v)=\frac{(C(s_i\cap t_v)-C(s_i\cap t_u\cap t_v))}{C(t_v)}\tag{9}$$

以上就是句子对齐过程中所采用的长度特征、对齐模式特征和共现汉字特征,利用统计模型将这 3 个特征组合起来,用动态规划的方法对整篇文章进行古文和现代文的句子对齐。

2　实验及结果分析

2.1　语　料

从《史记》全部 130 篇文章中挑选 12 篇作为实验对象[15-17]。对这 12 篇语料进行了人工句子对齐。从 12 篇中选出 8 篇作为训练语料,调整各特征的参数值,剩余 4 篇进行测试。训练语料和测试语料详见表 2。

表 2　训练、评测语料选用篇章

	序号	篇名	译者	古文句数	现代文句数
训练语料	卷七	项羽本纪第七	解惠全	595	595
	卷二十三	礼书第一	刘洪涛	154	170
	卷二十五	律书第三	刘洪涛	222	216
	卷二十六	历书第四	刘洪涛	61	64
	卷四十一	越王勾践世家第十一	黄永武	294	265
	卷四十八	陈涉世家第十八	宋尚斋	174	174
	卷一百七	魏其武安侯列传第四十七	宋尚斋	256	257
	卷一百十	匈奴列传第五十	王延海	474	467
总计				2230	2208
测试语料	卷二十九	河渠书第七	刘洪涛	98	92
	卷八十一	廉颇蔺相如列传第二十一	支菊生	239	237
	卷八十三	鲁仲连邹阳列传第二十三	王学孟	231	233
	卷一百二九	货殖列传第六十九	范君石	287	309
总计				855	871

2.2　评　测

假设机器对齐连接结果为 a,而正确的对齐连接结果为 a_r,A 为句子对齐过程中的正确对齐个数,B 为句子对齐过程中的错误对齐个数,C 为实验没有对齐但应该是正确的句子对齐个数,则召回率(R)、准确率(P)的定义为:

$$R=\frac{|a\cap a_r|}{|a_r|}=\frac{A}{A+C}\tag{10}$$

$$P=\frac{|a\cap a_r|}{|a|}=\frac{A}{A+B}\tag{11}$$

召回率和准确率都是衡量句子对齐的标准,均满足 $0\leqslant R\leqslant 1$,$0\leqslant P\leqslant 1$,召回率和准确率越高,说明句子对齐的效果越好。将召回率和准确率综合起来考虑的 F 为:

$$F=2\times\frac{R\times P}{R+P}\tag{12}$$

2.3　参数调整

整个程序中,一部分参数通过对语料的统计分析直接获得,而对数线性模型中各个特征的参数值则需要一步步尝试获得。如何获取对数线性模型中各个特征的参数值?目前常用的方法包括最大熵法、感知机、最小错误率学习法和互信息等。本文以最小错误率学习法思想为依据,以学习训练语料为目标,逐一调整特征参数值,即调整某一特征参数值的同时,固定其他参数值,使得学习语料准确率值达到最高。

2.4　实验结果

在实验过程中,选取不同的特征组合进行计算,得出以下实验结果如表 3 所示。

表 3　分别考虑不同特征或不同特征结合的对齐结果

	长度特征	对齐模式特征	共现汉字特征	R	P	F
1	√			84.3	50.3	63.0
2			√	88.8	51.5	65.2
3	√	√		93.7	94.1	93.9
4		√	√	90.5	91.8	91.2
5	√	√		93.7	94.1	93.9
6	√	√	√	94.3	94.4	94.3

从表 3 中可以看出,只考虑长度特征或共现汉字特征,准确率很低。若考虑三个特征中的两个特征,准确率、召回率和 F 值都较高。若同时考虑三个特征,准确率、召回率和 F 值是最高的。

2.5　实验结果分析

以句子对齐为例,结合实际语料,对不同特征组合的结果进行分析。

(1) 长度特征

如果只以长度特征进行句子对齐,计算机容易把两个 1-1 模式的正确句子对齐自动对齐成 2-2 模式,如表 4 所示,前两个对齐(1-1)为人工对齐结果,而(2-2)为机器自动对齐结果。

表 4　人工对齐与机器对齐结果比较

模式	古文	现代文
1-1	赵惠文王十六年,廉颇为赵将伐齐,大破之,取阳晋,拜为上卿,以勇气闻於诸侯。	赵惠文王十六年(前 283),廉颇率领赵军征讨齐国,大败齐军,夺取了阳晋,被封为上卿,他以勇气闻名于诸侯各国。
1-1	蔺相如者,赵人也,为赵宦者令缪贤舍人。	蔺相如是赵国人,是赵国宦者令缪贤家的门客。
2-2	赵惠文王十六年,廉颇为赵将伐齐,大破之,取阳晋,拜为上卿,以勇气闻於诸侯。 蔺相如者,赵人也,为赵宦者令缪贤舍人。	赵惠文王十六年(前 283),廉颇率领赵军征讨齐国,大败齐军,夺取了阳晋,被封为上卿,他以勇气闻名于诸侯各国。 蔺相如是赵国人,是赵国宦者令缪贤家的门客。

用图 1 表示人工对齐和机器对齐的结果,用动态规划算法的思路,人工对齐的路径为(0,0)→(1,1)→(2,2)→(3,3)→(4,4)→(5,5),而机器对齐结果则是(0,0)→(1,1)→(3,3)→(5,5)。我们针对人工对齐结果和机器对齐结果对各个点的路程值进行了计算,如果在只考虑长度特征的前提下,按照人工对齐结果中的(3,3)→(4,4)→(5,5)到达(5,5)的路程就会累积到 135,而通过 2 -2 模式直接(3,3)→(5,5),到达(5,5)的路程只有 118。可见,机器对齐会产生和人工对齐不一样的结果,从而使得准确率降低。

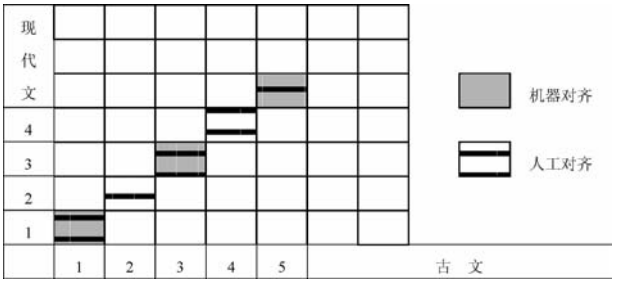

图 1 只应用长度特征情况分析

表 3 给出只利用句子长度特征对训练样本和测试样本进行句子对齐的结果,召回率较高,准确率较低,*F* 值较低。原因是本文所采用的召回率和准确率计算方法以人工对齐为依据,在人工对齐中主要对齐模式是 1 -1 模式,而机器对齐则把许多 1 -1对齐模式对齐成 2 -2 对齐,而 2 -2 对齐就会产生 4 个连接结果,这 4 个连接中有两个正确,两个错误。结果导致只应用长度特征进行句子对齐分析结果的召回率较高,而准确率较低。

(2) 共现汉字特征

如果只考虑共现汉字特征,往往会产生 $M \times N$ 模式对齐(M 个句子对齐 N 个句子)。因为考虑的句子个数多,共现汉字个数也多,所以容易产生多个句子对齐多个句子。这与语料中绝大多数句子对齐模式是 1 个句子对齐 1 个句子相矛盾。$M \times N$ 模式对齐同 2 -2 模式对齐同样原因使得结果召回率高,而准确率不高。

(3) 对齐模式特征 + 其他特征

若考虑共现汉字特征或长度特征外,还考虑对齐模式特征,则获得了较理想的结果。因为 1 -1 对齐模式概率高,而 2 -2 对齐模式和 $M \times N$ 对齐模式概率很小,结果使得考虑两个特征(共现汉字特征和对齐模式特征,长度特征和对齐模式特征)或三个特征产生的对齐多为 1 -1 对齐,而较少为 2 -2 对齐或 $M \times N$ 对齐。因此,正确率和召回率都较高。

3 结 语

本文就语句对齐这一方面对《史记》古文和现代文进行了初步的研究,应用了句长、对齐模式和共现汉字等特征,并取得了较为满意的结果。对于长度特征,采用以“字”作为古文和现代文句长的长度单位,而非“词”;对于共现汉字特征,也考虑“汉字”而非古文和现代文的互译词;对于对齐模式特征,我们只考虑 1 -1、1 -0、0 -1、1 -2、2 -1 和 2 -2 模式,而其他出现频率不高的模式没有在考虑的范围之内。

在以后的研究中,将进一步扩大研究语料,同时可以在古文人名、地名识别、特殊符号识别等方面展开进一步的研究。语料规模扩大和人名、地名等识别准确率提高必将促进句子对齐的准确率。

参 考 文 献

[1] Brown Peter F. Aligning Sentences in Parallel Corpora[C]//Proceedings of 29th Annual Conference of the Association for Computational Linguistics,1991:169 -176.

[2] William A Gale, Kenneth W Church. A Program for Aligning Sentences in Bilingual Corpora[C]//Proceedings of 29th Annual Conference of the Association for Computational Linguistics,1993:76 -101.

[3] Martin Kay, Martin Roscheisen. Text-translation Alignment[J]. Computational Linguistics,1993,19(1):121 -142.

[4] Michel Simard. Using cognates to align sentences in bilingual corpora [C]//Proceedings of the 4th International Conference on Theoretical and Mythological Issues in Machined Translation,1992:69 -83.

[5] Stanley F Chen. Aligning sentences in bilingual corpora using lexical information[C]//Proceedings of 31st Annual Meeting of the Association for Computational Linguistics,1993:9 -16.

[6] Wu Dekai. Aligning a parallel English-Chinese corpus statistically with lexical criteria[C]//Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics,1994:80 -87.

[7] Wang Xiaojie, Ren Fuji. Chinese-Japanese Clause Alignment[C]// Computational Linguistics and Intelligent Text Processing 6th International Conference,2005:400 -412.

[8] Krzysztof Jassem, Jaroslaw Lipski. A new tool for the bilingual text aligning at the sentence level[J]. Intelligent Information Systems,2008: 279 -286.

[9] Och Franz Josef, Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,2002:295 -302.

[10] Liu Yang, Liu Qun. Log-linear Models for Word Alignment[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,2005:459 -466.

[11] 王斌. 汉英双语语料库自动对齐研究[D]. 北京:中国科学院计算技术研究所,1999.

[12] 吕学强. 基于统计的汉英句子对齐研究[J]. 小型微型计算机系统,2004,25(6):990 -992.

[13] 林准. 古今汉语对齐研究[D]. 北京: 北京邮电大学信息工程学院,2007.

[14] 郭锐. 基于自动句对齐的相似古文句子检索[J]. 中文信息学报,2008,22(2):87 -92.

[15] 司马迁(汉). 史记[M]. 北京:中华书局,2006.

[16] 司马迁(汉). 白话史记[M]. 杨燕起, 等译. 长沙: 岳麓书社,2002.

[17] 司马迁(汉). 白话史记[M]. 台湾十四院校六十教授,合译. 北京: 新世界出版社,2007.

(上接第 94 页)

[9] Erchin Serpedin, Georgios B Giannakis. A Simple Proof of a Known Blind Channel Identifiability Result[J]. IEEE Trans on Signal Processing,1999,47(2):591 -593.

[10] Karakutuk S, Tuncer T E. Chnnel matrix recursion for blind effective channel order estimation[J]. IEEE Trans on Signal Processing,2011, 59(4):516 -525.

[11] Tong L, Xu G H, Kailath T. Blind Identification and Equalization Based on Second-order Statistics: A Time-domain Approach[J]. IEEE Trans on Inform Theory,1994,40(2): 340 -349.