汉语古现句子对齐研究

李闻
（清华大学，北京 100084）

摘要：本文关注汉语古现对齐方法，其中融入了新的特征，同时去掉了较常用的句长特征。通过句内分段将一对多、多对一和多对多模式中的长句进行拆分,使对齐的句对中不再有一对多、多对一和多对多模式。同时，对共现汉字特征和互信息特征进行了优化。从而使统计机器翻译中的词对齐、短语抽取，乃至整个统计机器翻译系统的质量水平都会有所提高。
关键词：句子；对齐；句内分段；拆分

Research on Alignment of Ancient Chinese Sentences to Modern Ones

Abstract: this paper focuses on method of sentence alignment of ancient Chinese to modern Chinese, where incorporates new features, meanwhile the common sentence length feature is canceled out, making the sentence pairs no one - many, many - one and many - many matches by segmenting within a sentence and spliting long sentences. Meanwhile, the issues on optimizing the features of co-occurred Chinese characters and mutual information are discussed. This will benefit the word alignment, phrase extraction, even the whole SMT system.
Key words: sentence; alignment; segmentation in a sentence; split

1 引言

统计机器翻译（SMT）从 20 世纪 90 年代初由 Brown（1993）成功实现至今，已走过了 20 余个年头。这期间，既有弯路，亦有高歌猛进。总体而言，统计机器翻译是比较成功的。目前，在诸多领域都可以见到它的身影，比如谷歌翻译、有道词典里的机器翻译等。既然是统计机器翻译，那就免不了要对语料进行统计，从中学习知识，然后才能对生句子进行翻译。这就需要大规模的双语平行语料库以让计算机从中充分学习知识。在跨语言检索中，双语平行语料库也是必不可少的。大规模双语平行语料库的创建用人工完成是不现实的，因为会耗费大量的人力和物力。再者，双语平行语料库是统计机器翻译的基础，它的质量高低直接影响到机器翻译的质量水平。所以建设双语平行语料库的句子对齐算法非常关键且一直倍受关注。

Brown 等（1991）首次提出了以词为单位的基于长度的句子对齐方法，他以英法语料为例进行了研究。英法两种语言属于形式上比较接近的语言，发现法英句长比为正态分布。对齐效果比较满意。Gale 和 Church（1993）也提出了基于长度的句子对齐方法，不过是以字符为单位，而非词。此方法更接近于汉字的模式。因为在计算共现汉字时一般也是以字为单位。Chen（1993）提出了基于词汇的对齐方法。该方法是目前比较有效的对齐方法之一，准确率较高。

基于长度的方法速度快，但准确率低，基于词汇的准确率低，但计算开销大。鉴于此，Li（2010）在 Ma（2006）所提出基于词汇的 Champollion 的基础之上，提出了句长与词汇相结合的 Fast-Champollion 算法，将句长信息与词汇信息有机结合，把输入的双语文本拆分为包含若干个句子的较小片段来进行对齐，其速度比 Champollion 算法提高了近 40 倍。

2 前人的工作

Wang 等（2005）对中日子句对齐方法进行了研究，采用三个特征。利用了句长信息、匹配模式和共现汉字信息。因为日语里有不少的汉字，所以共现汉字信息是一个比较好的特征。Liu 等（2012）对汉语古现句子对齐进行了研究和探索，也是采用了句长、匹配模式和共现汉字三个特征。

Gale 和 Church（1991）提到长句趋于翻译为长句，短句趋于翻译为短句。这种语言关系在大多数语言之间比较符合。Wang 等（2005）和 Liu 等（2012）分别统计计算出了日汉与汉语古现长度比，即 μ =1.47 和 σ =0.3 与 c ≈ 1. 617 和 s 2 ≈ 1. 56。然而，由于古汉语比较特殊，在句长方面，其与现代汉语的关系表现得并不是十分稳定，如表 1 中所示：

三月，匈奴入雁门。	三月，匈奴进入雁门。
以诸侯太盛，而错为之不以渐也。	这是由于诸侯王的势力太强大，而晁错在行动的时候不是采取渐渐削弱的方法。

表 1 古现句长比的不稳定性

表 1 中，第一行古汉语与现代汉语译文只差一个字，句长比只有 1.1 左右。而第二行古现句长比则超过 2.3。这种情况还比较普遍，所以基于句子长度的方法对于汉语古现对齐似乎不太适用。正如 Liu 等（2012）提到，计算机往往把两个 1-1 模式的正确对齐变为一个 2-2 模式的“错误”对齐（但从内容上讲是正确的），使平均句长增加一倍。这也反映出句长信息对于汉语古现对齐的不适用生。另外，Liu 等（2012）还提到，对于共现汉字特征，常常会出现多个句子对多个句子的 M-N 对齐情况，因为句子越多共现汉字也就越多。而这 M-N 对齐也许本应该是多个其他对齐，比如 1-1、1-2、2-1 等，所以我们称之为伪 M-N 对齐。如表 2 中所示。

伪 M-N 模式（7-4）	*又东三百里柢山*。**多水，无草木**。*有鱼焉，其状如牛，陵居，蛇尾有翼，其羽在魼下，其音如留牛，其名曰鰼，冬死而复生*。**食之无肿疾**。*又东三百里曰亶爰之山*。**多水，无草木，不可以上**。*有兽焉，其状如狸而有髦，其名曰类，自为牝牡，食者不妒*。	*再往东三百里，是座柢山，山间多水流，没有花草树木*。**有一种鱼，形状像牛，栖息在山坡上，长着蛇一样的尾巴并且有翅膀，而翅膀长在胁骨上，鸣叫的声音像犁牛，名称是鰼，冬天蛰伏而夏天复苏，吃了它的肉就能使人不患痈肿疾病**。*再往东三百里，是座亶爰山，山间多水流，没有花草树木，不能攀登上去*。**山中有一种野兽，形状像野猫却长着像人一样的长头发，名称是类，一身具有雄雌两种性器官，吃了它的肉就会使人不产生妒忌心**。

表 2　伪 **M-N** 模式

但在考虑对齐模式特征后，又可能会使 M-N 模式漏网，影响召回率和准确率等指标，因为只有 1-1 模式的概率比较高，其他模式都较低。另外，M-N 对齐句对中的一个句子可能由多个句子接合而成。多对多虽然从内容上讲没有错，但过长的句子对于以后的词对齐、短语抽取乃至整个统计机器翻译系统的质量水平都是不利的。可见源头错误对整个系统的影响之大。有鉴于此，我们提出了一个可以大大缩短汉语古现句子对齐中句子长度的方法，即不考虑句长和匹配模式特征，但考虑句内的分段情况。该方法不仅可以避免出现上述 Liu 等（2012）提到的 1-1 变 2-2 问题、伪 M-N 问题以及 M-N 模式漏网问题，而且对于真正的 M-N 对齐，也可以使对齐的句对中任一句子的长度不超过对齐前其语言的 M 或 N 个相关句子当中最长句子的长度。如式（1）所示。

LA <= max(L1, …, LM,N)　　(1)

其中，L 表示句长，A 表示对齐，M 和 N 分别为每种语言的句子的数目，且 M 与 N 不同时为 1。Li 表示第 i 个句子的长度。表 3 中显示了这种由多个句子接合成超长句子的真正 M-N 的情况。

2－2 模式	*又东五百里曰区吴之山*。**无草木，多沙石，鹿水出焉，而南流注于滂水**。	*再往东五百里，是座区吴山，山上没有花草树木，到处是沙子石头*。**鹿水从这座山发源，然后向南流入滂水**。
3－1 模式	*又东五百里曰浮玉之山*。**北望具区，东望诸　**。*有兽焉，其状如虎而牛尾，其音如吠犬，其名曰彘，是食人*。	再往东五百里，是座浮玉山，在山上向北可以望见具区泽，向东可以望见诸水，山中有一种野兽，形状像老虎却长着牛的尾巴，发出的叫声如同狗叫，名称是彘，是能吃人的。

表 3　真正的 **M-N** 模式

我们的截断分段方法客观上也是可行的。英语等语言的句子结构灵活而复杂，比如从句，插入语、同位语、时间和地点状态等。一般情况下，无法将句子真正断开。而汉语是顺序性的，没有那些复杂的结构。所以易于拆分开来。如表 9 中所示。

3 模型

由于对数线性模型的广泛适用性，很多关于词或句子对齐等研究中都选用了此模型。该模型从 Papineni（1997）提出至今，已在诸多领域成功应用，比如 Och 等（2002）用于统计机器翻译，Liu 等（2005）用于词对齐，以及 Liu 等（2012）用于句子对齐等。其特点是在特征数较多时可以随意融入特征，所有特征均被视为函数。该模型如式（2）所示：

$$Pr(\mathbf{a}|\mathbf{e},\mathbf{f}) = \frac{\exp[\sum_{m=1}^{M}\lambda_m h_m(\mathbf{a},\mathbf{e},\mathbf{f})]}{\sum_{\mathbf{a}'}\exp[\sum_{m=1}^{M}\lambda_m h_m(\mathbf{a}',\mathbf{e},\mathbf{f})]} \quad (2)$$

由于句长特征对汉语古现对齐的不适用性以及防止 m-n（m 和 n 分别为每种语言的句子的数目，且 m 与 n 不同时为 1）模式漏网，所以只考虑共现汉字特征和共现汉字间的互信息特征。这样就不应用此模型。我们采用基于句内层次分段的模型，首先一个句子可以分为若干个可独立成句的子句，表示如下：

1. 一对一的情况：

$$S = C \quad (3)$$

2. 一对多（1-n）的情况：

$$\begin{cases} Sg = C1 + \cdots + Cn \\ Sx = S1 + \cdots + Sn \end{cases}, \text{其中 Si = C} \quad (4)$$

3. 多对一（m-1）的情况：

$$\begin{cases} Sg = S1 + \cdots + Sm \\ Sx = C1 + \cdots + Cm \end{cases}, \text{其中 Si = C} \quad (5)$$

4. 多对多的情况：

$$S = C_1 + \cdots + C_{(NoP-1)} \quad (6)$$

其中，S 表示句子，C 表示子句，g 表示古汉语句，x 表示现代汉语，NoP 表示原始对齐的句对中源语言和目标语言句子内句号的总和。

对于 C：在多对多情况下，一般表示根据两种语言的句子中依次出现的句号进行相应拆分后得到的子句；在多对一或一对多的情况下，一般表示长句对应相对语言的短句进行拆分后得到的子句以及相对语言的短句；在一对一的情况下，C = S。

实际上，无论是一对一、一对多、多对一，还是多对多模式的原始对齐，也无论是源语言还是目标语言，用我们的方法最终得到的子句数均为 NoP － 1。

子句又可以再细分为分段 (Segment)，表示如下：

$$C = Seg_1 + Seg_2 + \cdots + Seg_k \quad (7)$$

Seg 表示句内或子句内的分段，即用逗号或分号分隔开的成分。那么有：

$$d(C_g, C_x) = H(C_g, C_x) + I(h_1, h_2, \cdots, h_k)$$

$$= H(Seg_{g1} + Seg_{g2} + \cdots + Seg_{gm}, Seg_{x1} + Seg_{x2} + \cdots + Seg_{xn}) + I(h_1, h_2, \cdots, h_k) \quad (8)$$

其中，H 是共现汉字信息，I 是汉字之间的互信息。

4 方法和实验

我们的方法不考虑句长特征和匹配模式特征，但考虑句内分段情况。虽然汉语古现句子的字数比极不稳定，但古现句内分段数之比是相当稳定的，变化很小。这样，我们就可以利用这一特点来实现古现句子的对齐。

4.1 共现汉字特征优化

对于共现汉字，有时由于共现的字无法满足需要，使得对齐效果不好。考虑到这一点，我们通过统计方法找出古汉语中的一些重要的字，并考察其在现代汉语中的译法而制作了一个简明字典，以增加共现汉字的个数。如图 1 和图 2 所示：

言	说
欲	想
一	
事	
可	
卒	死\去世\最终\终于\最后\
汉	
皆	全\都
东	

图 1　古汉语中关键字词的选取

之	的\他们\他\她\它\其
曰	说\叫\是
二	两
至	到
无	没有
言	说
欲	想
卒	死\去世\最终\终于\最后
皆	全\都
矣	了
闻	听
遂	于是\就\便\终于
吾	我
余	多
日	天\太阳
亦	也
岁	年
诛	杀\处死\讨伐
耳	罢了/吧
若	像\如果
首	先\头
尽	全\都
弗	不\没有\无
斩	杀

图 2　简明词典

这个词典与普通词典的不同之处在于，只包含为数不多的一些关键的字，如图 2 中所示。这样，对处理速度影响不大。虽然该简明字典中的条目有限，但其贡献较大，因为都是最常见的一些字且在现代文中的表达发生了较大变化。对于古汉语句子中没有匹配上的字，到简明辞典里查一下，然后用辞典里的字或二字词再重新匹配。调整后古汉语句子中的一个字可能变为两个字，此时如果这两个字与现代汉语中的两个字匹配上，也算古汉语句中一个字匹配上，即算一个共现汉

字，不影响互信息的计算。

4.2 互信息特征优化

对于互信息特征，由于古汉语与现代汉语之间的差异，使得古汉语中本来相邻的两个字在现代汉语中分开了。为了解决这一问题，提出了以下三种解决方案。

4.2.1 可去掉的字

在现代汉语中，像“了”、“的”这样的字没有实际意义，可以去掉而不会影响语义。同样，量词也可以去掉。所以在匹配时忽略“了”、“的”以及量词，以便更多地利用互信息。如表 4 中的斜体字，去掉之后互信息就会得到更加充分的利用。

是岁，越王句践灭吴。	这一年，越王句践灭*了*吴国。
毋恤群臣请死之。	随从毋恤*的*群臣要求把知伯处死。
其後娶空同氏，生五子。	后来襄子娶空同氏为妻，生*了*五*个*儿子。

表 4 现代汉语中可以去掉的字

4.2.2 单双字词问题

古汉语一般是单字词，而现代汉语则一般为双字词。根据这一特点，在匹配时可以将现代汉语以两个字为一个单元进行匹配，如果现代汉语两字单元中的任何一个字匹配上某一古汉语字，则认为该古汉语字与现代汉语两字单元匹配。可以称之为“字-单元匹配法”。例如，表 5 第一行古汉语句子中的“怨”与现代汉语中的二字单元“怨恨”匹配；第二行古汉语句子中的“草”、“木”、“ 沙”和“石”分别与现代汉语句子中的二字单元匹配。这样就达到了充分利用互信息的目的。

毋恤由此怨知伯。	毋恤从此更加怨*恨*知柏。
无草木，多沙石	山上没有*花*草*树*木，到处是沙*子*石*头*。

表 5 单双字词问题

4.2.3 排列组合问题

通过上述两种方法，显然可以更充分地利用互信息。但是，有时古现汉语句子中词或词组的顺序是不同的（如表 6 中的“县五十二”对应的“五十二个县”），或者由于增译而将原本相邻的词或词组分隔开（如表 6 中的“三百石十一人”对应的“三百石的官员十一人”）。此时，可以考虑用组合而非排列方法来充分利用互信息特征，但组合的范围应在一个分段内。例如，表 6 中，分段“破军七”在与分段“打垮过七支敌军”进行匹配时，搜索范围不能超越后者。在分段“打垮过七支敌军”中对“军”进行搜索时发现匹配项；在对“七”进行搜索时又在其中发现匹配项。那么就可以认为在现代汉语句子分段“打垮过七支敌军”中，“军”与“七”是顺序相邻的。同理，在现代汉语句子分段“二千石以下到三百石的官员十一人”中，认为“三百石”与“十一人”是相邻的，可以连接成一个字符串。这样，就可以充分利用互信息特征。

别，破军七，下城五，定郡六，县五十二，得丞相一人，将军十二人，二千石已下至三百石十一人。	他自己单独带兵打仗，打垮过*七*支敌*军*，攻下过*五*座*城*池，平*定*了*六*个*郡*，*五十二*个*县*，并俘虏过敌人丞相一人，将军十二人，二千石以下到*三百石*的官员*十一人*。

表 6 排列组合问题

根据我们对《山海经》、《孙子兵法》和《诗经》等 5 篇语料标注的 Bead 模式信息，统计出了关于 Bead 的数据，如表 7 中所示：

Bead 类型（m-n）	数量	Pr (%)	总计
0-1,1-0	14	3.3254157	421
1--1	325	77.19715	
1-2,2-1	75	17.814727	
1-3,3-1	2	0.4750594	
1-4,4-1	1	0.2375297	
2--2	4	0.9501188	

表 7 Bead 类型及分布

从中，可以看出一对多或多对一的情况是比较普遍的，而多对多，只发现了 2-2 的情况。但是，我们的算法对于一对多、多对一以及多对多匹配模式，会全部考虑。这样，基本不会出现上文提到的 M-N 模式漏网的情况。

对于表 3 中的第二行，进行句内分段匹配时，会出现误配的情况，如表 8 所示。

	又东五百里曰浮玉之山。	再往东五百里。
	北望具区，东望诸 。	是座浮玉山，在山上向北可以望见具区泽，向东可以望见诸水。

表 8 错配

这是因为〝又东五百里曰浮玉之山〞在译为现代汉语后被拆分成了两个段，即“再往东五百里”和“是座浮玉山”。〝又东五百里曰浮玉之山〞在与〝再往东五百里，是座区吴山，山上没有花草树木，到处是沙子石头。〞内的各个分段进行尾

段匹配时，其与"再往东五百里"的共现汉字个数要多于其与"是座浮玉山"的共现汉字个数。所以会认为“再往东五百里”是"又东五百里曰浮玉之山"对应译文的尾段。

这里的尾段匹配法是指在匹配两个句子时，首先计算出各自的分段数，然后将分段数较少的句子的尾段与分段数较多的句子中的各个分段依次进行比较，按共现汉字和互信息计算并记录比较的得分。最后，取分段数较多的句子中得分最高的分段作为该语言的子句尾。

解决上述问题的方法其一是在进行句内分段匹配时，记录源语言中与目标语言分段匹配上的最后一个字的位置，然后将其后面的字与目标语言中后边紧邻的一个分段进行匹配，如果其后面的字与目标语言相邻段的匹配率达到某一阈值，则将目标语言的这个相邻分段与其前面的相关分段接合为一个分段并与源语言子句对齐。另一种方法是根据现代汉语句子长度一般不会小于古汉语句子长度这一特点，在匹配前，先检查现代汉语句子长度是否大于古汉语句子长度，如果不大于则将现代汉语句子中相邻的两个相关段合并，然后再检查，依此类推。分段接合达到长度要求后再进行匹配。

1－1/2 模式	*又东五百里曰区吴之山。*	*再往东五百里，是座区吴山。*
1/2－1/2 模式	**无草木，多沙石。**	*山上没有花草树木，到处是沙子石头。*
1/2－1 模式	**鹿水出焉，而南流注于滂水。**	**鹿水从这座山发源，然后向南流入滂水。**
1－1/3 模式	*又东五百里曰浮玉之山。*	再往东五百里，是座浮玉山。
1－1/3 模式	**北望具区，东望诸　。**	在山上向北可以望见具区泽，向东可以望见诸水。
1－1/3 模式	*有兽焉，其状如虎而牛尾，其音如吠犬，其名曰彘，是食人。*	山中有一种野兽，形状像老虎却长着牛的尾巴，发出的叫声如同狗叫，名称是彘，是能吃人的。

表 9　正确的配对

对于上文表 2 中的例子，表 9 中显示了用我们的方法对齐后的结果。对于句子拆分的情况，我们在表 9 中用分数的形式表示，如 1/2 表示拆分为 2 个句子，1/3 表示拆分为 3 个句子，等等。

对于 1 – M、M – 1 或 M – N 模式，用我们的方法得到的对齐结果中，无论是源语言还是目标语言，最大句子长度等于对齐前其语言的 M 或 N 个相关句子当中最长句子的长度。这是做也是合理的，因为既然源或目标语言的句子的一部分可以在目标或源语言中用一个完整的句子表示，那么这一部分就应该可以分离出来单独成句，而基本不会影响整体语义。

根据汉语有顺序性的特点，对于一对多或多对一的情形，一般是把逗号变为了句号，而不会把句子从没有标点符号的地方截断，因而句内分段点不会改变。或者说古汉语的断句和现代汉语的翻译都有合理之处，在处理古现多对一的情况时，我们采用古汉语的断句，如表 11，而在处理古现一对多的情况时，我们采用现代汉语译文的断句如表 10 和表 12。

太史公曰：汉兴，孝文施大德，天下怀安，至孝景，不复忧异姓，而晁错刻削诸侯，遂使七国俱起，合从而西乡，以诸侯太盛，而错为之不以渐也。	太史公说：汉兴以来，孝文皇帝广施大德，天下百姓怀恩而安。
	到了孝景皇帝，不再忧虑异姓诸侯王。
	然而晁错削夺同姓诸侯王的封地，导致七国共同起兵，联合西向。
	这是由于诸侯王的势力太强大，而晁错在行动的时候不是采取渐渐削弱的方法。

表 10　选择现代汉语中的断句

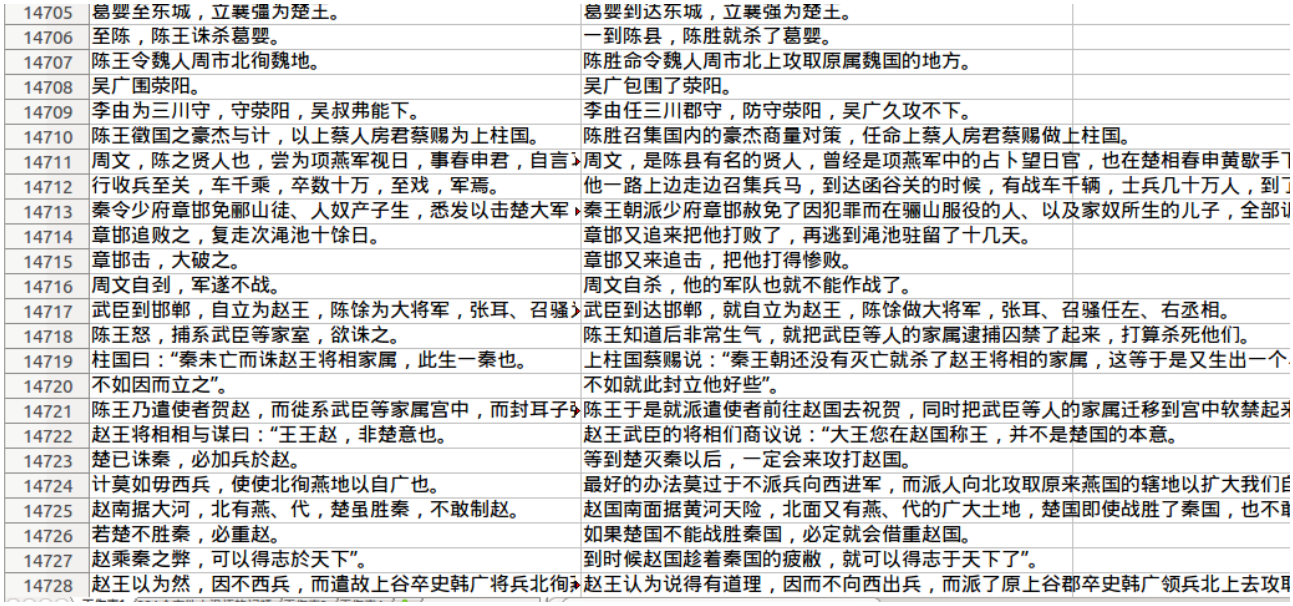
英水出焉，西南流注于赤水。	英水从这座山发源，向西南流入赤水，水中有很多白色玉石，还有很多粟粒般大小的丹沙。
其中多白玉，多丹粟。	

表 11　选择古代汉语中的断句

中元年，封故御史大夫周苛孙平为绳侯，故御史大夫周昌左车为安阳侯，四月乙巳，赦天下，赐爵一级。	中元元年，封原来的御史大夫周苛的孙子周平为绳侯，原来的御史大夫周昌的孙子周左车为安阳侯。
	四月二十三日，大赦天下，赐予百姓爵位一级。

表 12　选择现代汉语中的断句

图 3 是对齐结果示例。



14705	葛婴至东城，立襄彊为楚王。	葛婴到达东城，立襄强为楚王。
14706	至陈，陈王诛杀葛婴。	一到陈县，陈胜就杀了葛婴。
14707	陈王令魏人周市北徇魏地。	陈胜命令魏人周市北上攻取原属魏国的地方。
14708	吴广围荥阳。	吴广包围了荥阳。
14709	李由为三川守，守荥阳，吴叔弗能下。	李由任三川郡守，防守荥阳，吴广久攻不下。
14710	陈王徵国之豪杰与计，以上蔡人房君蔡赐为上柱国。	陈胜召集国内的豪杰商量对策，任命上蔡人房君蔡赐做上柱国。
14711	周文，陈之贤人也，尝为项燕军视日，事春申君，自言	周文，是陈县有名的贤人，曾经是项燕军中的占卜望日官，也在楚相春申黄歇手
14712	行收兵至关，车千乘，卒数十万，至戏，军焉。	他一路上边走边召集兵马，到达函谷关的时候，有战车千辆，士兵几十万人，到
14713	秦令少府章邯免郦山徒、人奴产子生，悉发以击楚大军	秦王朝派少府章邯赦免了因犯罪而在骊山服役的人、以及家奴所生的儿子，全部
14714	章邯追败之，复走次渑池十馀日。	章邯又追来把他打败了，再逃到渑池驻留了十几天。
14715	章邯击，大破之。	章邯又来追击，把他打得惨败。
14716	周文自刭，军遂不战。	周文自杀，他的军队也就不能作战了。
14717	武臣到邯郸，自立为赵王，陈馀为大将军，张耳、召骚	武臣到达邯郸，就自立为赵王，陈馀做大将军，张耳、召骚任左、右丞相。
14718	陈王怒，捕系武臣等家室，欲诛之。	陈王知道后非常生气，就把武臣等人的家属逮捕囚禁了起来，打算杀死他们。
14719	柱国曰："秦未亡而诛赵王将相家属，此生一秦也。	上柱国蔡赐说："秦王朝还没有灭亡就杀了赵王将相的家属，这等于是又生出一个
14720	不如因而立之"。	不如就此封立他好些"。
14721	陈王乃遣使者贺赵，而徙系武臣等家属宫中，而封耳子	陈王于是就派遣使者前往赵国去祝贺，同时把武臣等人的家属迁移到宫中软禁起来
14722	赵王将相相与谋曰："王王赵，非楚意也。	赵王武臣的将相们商议说："大王您在赵国称王，并不是楚国的本意。
14723	楚已诛秦，必加兵於赵。	等到楚灭秦以后，一定会来攻打赵国。
14724	计莫如毋西兵，使使北徇燕地以自广也。	最好的办法莫过于不派兵向西进军，而派人向北攻取原来燕国的辖地以扩大我们自
14725	赵南据大河，北有燕、代，楚虽胜秦，不敢制赵。	赵国南面据黄河天险，北面又有燕、代的广大土地，楚国即使战胜了秦国，也不敢
14726	若楚不胜秦，必重赵。	如果楚国不能战胜秦国，必定就会借重赵国。
14727	赵乘秦之弊，可以得志於天下"。	到时候赵国趁着秦国的疲敝，就可以得志于天下了"。
14728	赵王以为然，因不西兵，而遣故上谷卒史韩广将兵北徇	赵王认为说得有道理，因而不向西出兵，而派了原上谷郡卒史韩广领兵北上去攻取

图 3　对齐结果片断

5 评测

句子对齐的质量评测由式（4）-（6）算得：

$$Precision = \frac{|GB \cap PB|}{|PB|} \tag{9}$$

$$Recall = \frac{|GB \cap PB|}{|GB|} \tag{10}$$

$$F1\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

其中，GB 是人工标注的 Bead 集，PB 是我们的算法产生的 Bead 集。表 13 是评测的结果。

Precision	94.1%
Recall	93.8%
F1-measure	93.9%

表 13　评测结果

6 总结

本文提出的新特征句内分段对汉语古现对齐的效果是显著的。大大缩短了句对中句子的长度。为汉语古现机器翻译的后续的各项工作打下了其实的基础。不过本文的分析方法还不够全面，将来的工作是通过引入更多的特征来提高句子对齐质量，使句子对齐更上一层楼，且不留死角。

参考文献

1. Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. The mathematics of statistical Machine Translation: Parameter estimation[J]. Computational Linguistics 19, 2, 1993: pp.263-311.

2. Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora[C]. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, California, USA: Association for Computational Linguistics, 1991: pp.169- 176.

3. William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora[C]. In Proc. of 29th Conf. of Assoc. for Comput. Linguistics. ACL-1991: pp.177-184.

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

5. Peng Li and Maosong Sun. Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm[C]. Coling 2010: Poster Volume. 2010: pp.710-718.

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

7. Ying Liu, Nan Wang. Sentence Alignment for Ancient-Modern Chinese Parallel Corpus[C]. The 2012 International Conference on Web Information Systems and Mining and the 2012 International Conference on Artificial Intelligence and Computational Intelligence. 2012: pp.408-415.

8. Kishore A. Papineni, Salim Roukos, Todd Ward. Feature-based language understanding[C]. In Eu- ropean Conf. on Speech Communication and Technology. Rhodes, Greece,1997: pp.1435-1438.

9. Franz J. Och, Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation[C]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, PA, 2002: pp.295-302.

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

作者联系方式：
姓名：李闻
地址：北京海淀区清华园一号文北楼
邮编：100084
电话：15652503601
电子邮箱：Levin2@126.com