融合词对齐位置映射关系的句对齐算法*

尹宝生，王伟，吴闯，叶娜，蔡东风

（沈阳航空航天大学 人机智能研究中心，辽宁 沈阳市 110134）

摘要：本文利用双语文本中的词语位置关系，提出了融合词对齐位置映射关系的句对齐算法。该方法与传统词汇方法相比，不完全依赖于词典匹配结果，还通过词语的位置映射关系对匹配词对的位置合理性进行判断，同时对于未匹配的词语，根据其对应位置关系，给予一定的对齐概率。在此基础上，结合基于词汇字节长度的锚点分割策略，显著提高了算法的运行效率。在多类中英双语语料对齐的实验上，本文提出的方法有效提高了句子对齐效果。相比目前较健壮的 Champollion 句子对齐工具，在句子对齐的错误率上平均下降了 27.5%。

关键词：锚点分割策略；词对齐位置映射关系；句子对齐

中图分类号：TP391　　**文献标识码：**A

Incorporating Position Mapping Relationships from Word Alignment into Sentence Alignment Algorithm

YIN Baosheng, WANG Wei, WU Chuang, YE Na, CAI Dongfeng

(Human-Computer Intelligence Research Center, Shenyang Aerospace University,
Shenyang, Liaoning 110136, China)

Abstract: This paper puts forward a sentence alignment algorithm using the word position relationships in bilingual texts. Compared with the traditional lexicon-based algorithm, this algorithm does not entirely depend on words matched through the dictionary, but also judges the rationality of positions of matched word pairs through the position mapping relationships among words. The mismatched words are also taken into account and given a certain alignment probability through the corresponding position information. On this basis, the execution efficiency of the algorithm is increased greatly in combination with the efficient anchor-based segmentation strategy. Experimental results on Chinese-English bilingual corpora show that the proposed algorithm greatly improves the performance of sentence alignment. Compared to the state-of-the-art Champollion sentence alignment tool, the error rate decreases by 27.5%.

Key words: Anchor-based Segmentation Strategy; Position Mapping Relationships from Word Alignment; Sentence Alignment

1　引言

双语句对齐技术一直在机器翻译领域中扮演着极其重要的角色。同时也是使双语语料库实用化的关键基础步骤。在自然语言处理领域具有重大研究意义[1-2]。双语语料库的对齐形式有多种，如篇章级对齐、句子级对齐、词语级对齐，其中篇章级对齐较为容易，从篇章一级的双语文本中自动获得句子级对齐是双语句子对齐的主要过程，同时

* **收稿日期：**　　**定稿日期：**

基金项目：辽宁“百千万人才工程项目”（04021401）；国家自然科学基金（61402299）

作者简介：尹宝生（1975—），男，副教授，主要研究方向：知识管理和机器翻译；王伟（1990—），男，硕士研究生，主要研究方向：自然语言处理，机器翻译；吴闯（1985—），男，硕士，主要研究方向：自然语言处理，机器翻译；叶娜（1981—），女，讲师，博士，主要研究方向：自然语言处理，机器翻译；蔡东风（1958—），男，博士，博士生导师，教授，主要研究方向：人工智能。

也是进行下一步词语对齐工作的基础。

国内外学者早已把双语语料句子级自动对齐算法作为一个重要的研究课题并进行了大量的研究工作。目前句子对齐的算法主要分三类：第一类是基于长度的方法，Brown[3]和Gales[4]等人最先尝试利用长度信息的方法进行双语文本对齐，其主要思想是一种语言中长句翻译成另外一种语言后句子仍然是较长的，反之，短句翻译成另外一种语言后句子也仍是较短的。吕学强等人[5]、张霞等人[6]也分别对基于长度方法的评价函数和长度计算方法进行了相关研究。该类方法虽然对齐速度很快，但容易造成错误的蔓延，且在噪声较大的语料中对齐准确率较低；另外两类是基于词汇信息的方法和基于长度和词汇信息混合方法，这两类方法也是目前研究较为主流的方法，Chen[7]通过构建一种词汇到词汇的翻译模型，在英法语句之间实现了一种基于词汇的对齐方法。Wu[8]利用基于长度与词汇方法相结合的方式对汉英双语语料进行对齐,准确率达了 92.1%。钱丽萍、赵铁军[9]等人利用英汉互译文本之间的内在联系，通过使用一部较完整的词典作为桥梁，根据评价函数以及动态规划算法实现了汉英语句对齐，但其对齐效果对词典规模的依赖性较强。Ma[10]提出利用信息检索中的tf-idf思想给匹配到的词语赋予权重，实现了一种基于词汇信息的句子对齐算法，并在噪音高的语料库中表现出很高的准确率，但对齐速度较慢。Trieu[11]等人利用词聚类方法对Moore[12]的句对齐方法进行了改进。Li[13]等人和 MaimaitiminS[14]等人提出基于锚点句对的方法对文本进行片段分割，再在片段内进行对齐,该类方法有效的解决了大规模句子对齐任务中错误蔓延的问题，并大大提高了算法效率。还有一些学者利用机器学习算法等较新颖的方式进行句子对齐。Ru[15]等人、Mohamed Abdel Fattah[16]等人利用支持向量机（SVM）和神经网络（Neural Networks）等机器学习算法进行了句子对齐算法的研究。Ghaly[17]提出一种基于几何学方法实现的句子对齐系统，利用词汇线索，针对 English-Arabic 语料中的复杂对齐类型的句对进行对齐。这类方法的尝试，只是验证了一些机器学习方法在句子对齐任务上的有效性，但对句子对齐任务本身并没有突出贡献，对齐效果提升也不明显，同时大多文章只是针对自身语料的对齐难点进行了改进提高，所以该类方法的普适性较差。

综合上述方法的研究与分析，本文提出了融合词对齐位置映射关系的句子对齐算法。该句子对齐算法主要有两点贡献。第一点，引入词对齐位置映射关系这一新特征，与传统基于词汇的方法相比，不完全依赖于词典信息，还借助匹配词对的对应位置信息，对其进行匹配合理性判断，即匹配位置越合理的词对，其重要程度越大，同时对未匹配的词对，也通过其对应位置，给予一定的对齐概率。这种方式更加合理的分配匹配词对的权重，从而有效提高了句子互译相似度的计算效果。并且该特征不受语言种类的限制，可以融入到任何句子对齐方法中。第二点，提出高召回率的锚点识别方法，结合基于锚点的分割策略，显著提高算法的运算速度。

2 算法框架

本文采用基于锚点和词汇信息相结合的对齐方法，在中英双语语料上进行了句子对齐研究。首先利用单纯基于长度的句对齐方法对双语文本进行初对齐，获得候选锚点集，然后利用基于词汇字节长度的锚点识别方法对候选集进行筛选，得到最终锚点句对集合，再利用锚点句对将双语文本分割成相对应的多个片段，在片段内使用融合词对齐位置映射关系的句对相似度计算方法进行句子对齐。算法整体框架流程如图 1。

2．1 基于词汇字节长度的锚点识别

无论是对大规模双语语料库的对齐，还是小规模的双语文本的对齐，利用锚点对语料进行切分的方式，都是一个杜绝对齐错误蔓延、有效提高算法效率的手段。锚点句对识别结果的好坏直接影响最终对齐的效果，如果锚点句对识别错误，则会使双语本文片

段切分错误，导致后续段内对齐产生更多的错误，所以锚点句对必须为完全正确的互译句对。

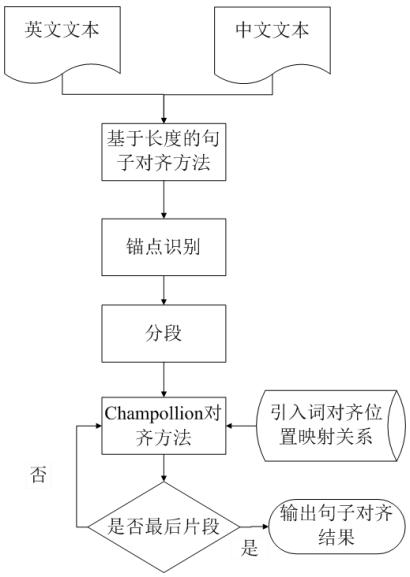

图 1 算法流程图

Li (2010) 等人利用锚点句对切分文本的方式显著提高了句子对齐算法的效率，并提出“指纹信息”来筛选锚点句对，主要思想是利用当前句子中与前一句和后一句中不同的词语作为自身的“指纹信息”，然后利用双语字典进行匹配。由于该方法不仅要提取“指纹信息”，而且还要进行中英和英中双向的匹配度计算，使得锚点句对的识别时间复杂度较大，召回率较低。

本文单纯利用匹配文本长度信息占双语句子总长度的比例来筛选锚点句对。该方法只进行英到汉的单向匹配，大大提高了锚点句对识别的计算速度。传统计算双语句子的匹配度往往需要对中文进行分词操作，但由于分词处理导致的错误，使得双语句对的词语匹配率有所下降，本文从三类领域中各抽取 10000 双语句对，进行匹配缩减率的统计，结果如表 1 所示：

表 1 分词导致的匹配缩减率

语料	匹配缩减率
HK 会议记录	24.97%
HK 法律	21.24%
HK 新闻	25.22%
平均	23.81%

从表 1 看出，分词后导致约 24%的词汇无法匹配。所以本文锚点句对识别过程中不对中文进行分词处理。如果英文句子中一个单词的译文或其本身在某个中文句子中出现，则认为这个英文单词为英文句子中匹配成功的文本，对应出现在中文句子中的译文为中文句子中匹配成功的文本，具体的评价函数如公式（1）所示：

$$MD(E,C)=\frac{Len(MatchT(E))+Len(MatchT(C))}{Len(E)+Len(C)} \quad (1)$$

公式(1)中，E、C 分别表示英文句子和中文句子。MatchT (E)表示匹配成功的英文文本，MatchT (C)表示匹配成功的中文文本，Len()方法表示计算文本的长度方法。从公式（1）可以看出，匹配到的英中文本长度占英中句子总长度的比值越大，则匹配度越高，说明两个句子成为锚点句对的可能性越大。但在实验中发现，不同的长度计算方式，对锚点句对识别的评价函数影响很大，合理有效的长度计算方式，也是使评价函数达到最佳效果的关键。所以针对文本长度的度量方式，本文进行了讨论研究，并尝试了三种不同的长度计算方式。

1. 将英汉文本的长度统一换算到字节上，具有格式统一、统计方便的特点。方法命名为 match-byte。

2. 考虑英汉各自的语言特点，英文文本以单词个数为基本长度单位，而中文文本以文字个数为基本长度单位。方法命名为 match-word。

3. 考虑到字典规模的限制，大量英文单词尤其是人名、地名、缩略语等无法在词典中找到，使匹配度急剧下降，针对此类情况熊伟[18]等人在考虑英汉各自语言特点的基础上,提出补偿方案，具体做法是，当英文单词在词典中不存在时，令英文句子长度减 1，汉语句子长度减 2（一个英文单词（人名）对应 2 个汉字）。方法命名为 match-reduce。

针对上述三种词汇长度计算方式，本文在三类数据集上进行了实验，具体实验数据见第三章。

2．2 融合词汇位置映射关系的句对相似度计算方法

句子对齐任务中的基本问题就是计算源语言文本和目标语言文本的相似程度。为使算法健壮性更强，在噪声语料和规范语料中都能有较好的表现，本文比较了champollion，hunalign 等较成熟的句子对齐算法，最终选择在多类数据集中都表现较好的 champollion 算法作为基础算法。其主要思想是借助信息检索中 tf-idf 权重计算方式来衡量匹配词语对的重要性，即出现频率越少的翻译对,其重要程度越高。

在此基础上，本文引入了词对齐位置映射关系，对 champollion 方法中的句对相似度计算公式进行了改进。现有的基于词汇的句子对齐算法，完全依赖词典进行词汇匹配，对匹配到的词汇并没有考虑其位置的合理性，即词汇在双语句对中匹配成功，该匹配词对就是正确的。在不互为译文的句对中，也有很多匹配成功的词对，但是其中却有很多匹配词对是不合理的。本文假设正确的匹配词对，其对应位置映射关系越强。即匹配词对中,位置越合理的,其重要度越高，而对于位置不合理的词对会给予相应的惩罚。本文使用基于统计方法的词对齐工具 GIZA++(https://github.com/alexey-osipenko/giza-pp)对十万句对英汉对齐双语语语料进行词对齐训练，获得词语位置映射关系信息，统计出各类对齐位置关系的频率，并将其转化为相应概率。以此对匹配词对的位置合理性进行判别。由于本文使用的词典规模较小，大量词汇无法通过词典得到匹配。针对未匹配的词语，本文同样利用其对应位置关系，给予一定的匹配概率。

定义 stf 为某词语在句子中的频率，即某词语在该句子中出现的次数,idtf 为某词语在整个文档中的逆词汇频率，具体计算如公式(2)所示：

$$idtf = \frac{T}{occurences\_in\_document} \quad (2)$$

T 表示整个文档中所有词语的个数，occurences_in_document 表示某个词语在文档中出现的个数。

假设 E,C 分别表示英文句子和中文句子，其具体表示形式如下：

$$E=\{e_1,e_2,.....,e_{m-1},e_m\}$$
$$C=\{c_1,c_2,.....,c_{n-1},c_n\}$$

e_i 和 c_j 分别表示分词后的单词，假设中文句子和英文句子中有 k 对匹配词对，其用集合 P 表示，未匹配的中文词语用集合 UC 表示，未匹配的英文词语用集合 UE 表示，公式如下：

$$P=\{(e_1^{'},c_1^{'}),(e_2^{'},c_2^{'})...(e_k^{'},c_k^{'})\}$$
$$UC=\{c_1^{''},c_2^{''}...c_{uc\_n}^{''}\}$$
$$UE=\{e_1^{''},e_2^{''}...e_{ue\_n}^{''}\}$$

词汇 $(e_1^{'},c_1^{'})$ 的位置合理性打分函数的计算公式如下：

$$AlignS(e_1^{'},c_1^{'})=log(1+p(site(e_1^{'}),site(c_1^{'}))) \quad (3)$$

其中 $p(site(e_1^{'}),site(c_1^{'}))$ 表示词语对 $(e_1^{'},c_1^{'})$ 位置对齐概率，通过上述 GIZA++词对齐工具训练得到。$site(e_1^{'})$ 表示计算单词 $e_1^{'}$ 在句子中的位置。

则最终融合位置映射关系的相似度计算方法，具体定义如公式(4)：

$$Sim(E,C)=(\sum_{i=1}^{k}(log(stf(e_i^{'},c_i^{'})*idtf(e_i^{'}))$$
$$+W_{matching}*AlignS(e_1^{'},c_1^{'}))$$
$$+W_{mismatching}*\frac{1}{ue\_n*ue\_n}\sum_{i=1}^{ue\_n}\sum_{j=1}^{uc\_n}AlignS(e_i^{''},c_j^{''}))$$
$$*alignment\_penalty_{ij}*length\_penalty(E,C) \quad (4)$$

上述公式中 $alignment\_penalty_{ij}$ 参数和 $length\_penalty(E,C)$ 的计算方式完全按照 champollion 算法中的计算方式计算。$W_{matching}$ 表示匹配词对位置对齐概率的权重，$W_{mismatching}$ 表示未匹配词对位置对齐概率的权重。

3 实验与分析

为了验证本文提出的锚点识别方法的

有效性，分别从 HK 新闻、HK 会议记录和 HK 法律三类已经句子级对齐好的双语语料中各抽取 5000 句正确句对，作为实验的正例，负例通过将 5000 句正确句对中的中文句子用邻近的前三个句或后三句中的任意一句替换的方式产生。在三类测试数据集上分别与 Li(2010)等的锚点识别方法进行了比较。由于每种方法在三类数据集上取得最好效果的阈值各不相同，同时考虑到锚点识别评价函数的普适性，故选择在三类数据集上的最高阈值作为每种长度计算方法的最终阈值。实验结果如表 2 所示：

表 2 锚点识别结果

锚点识别方法	阈值	锚点个数		
		新闻	会议记录	法律
Li(2010)	0.58	5	13	73
match-byte	0.73	77	166	410
match-word	0.6	47	72	188
match-reduce	0.75	4	0	92

从表 2 可以看出，利用字节个数计算文本长度的方式，其锚点识别效果最好。针对 match-byte 方法，本文详细分析了未分词处理和以字节个数计算词汇长度这两种处理方式对该方法的影响。实验结果如表 3：

表 3 分词处理与长度计算方式分析结果

长度计算方式	阈值	锚点个数		
		新闻	会议记录	法律
词语个数（分词）	0.75	6	20	159
词语个数（未分词）	0.75	24	48	223
字节个数（分词）	0.7	50	67	215

为了验证本文提出的词对齐位置映射关系对句子对齐算法的有效性，本文采用 Champollion 句子对齐工具中自带的评测语料作为实验的测试集，（下载地址 http://champollion.sourceforge.net），其中包含新闻，法律，会议记录等多个领域。针对测试集中出现的各类对齐类型进行了频率统计，统计结果如表 4 所示。

本文实现的句子对齐算法采用动态规划策略实现句子的对齐过程.算法中只考虑了（1:1）、（0:1）、（1:0）、（2:1）、（1:2）、（2:2）六种对齐形式。实验中，中英双语词表和英文停用词表均使用 champollion 工具中自带的资源，本文只使用英中词典,经去停用词处理后,共 4807 个英文单词,英文采用 ntlk（http://www.ntlk.org）工具包中的分词工具进行英文分词，中文采用 champollion 工具中的中文分词工具，并进行语料预处理，例如标点符号、汉语数字、特殊符号等转换。词对齐训练语料采用 NiuTrans 系统中的十万句中英双语语料。首先从三类语料中各选取一个测试文档作为开发集，在开发集上枚举 $W_{matching}$ 和 $W_{mismatching}$ 的组合可能，选择在多个数据集上都使算法显著提升的参数值最为最终参数。同时对比了单独引入匹配成功词对的位置映射关系和单独引入未匹配词对的位置映射关系对句子对齐算法的影响，结果如图 2。

最后通过人工经验的方式，$W_{mismatching}$ 值取 120，$W_{matching}$ 值取 140。 在 3 类语料的 3 个测试集上的实验结果如表 5 所示。

针对基于锚点对齐策略对句子对齐算法对齐效率的影响，本文通过结合法律、新闻、会议记录三类语料，形成中英句子数各约 2000 句的待对齐文本，统计了不同对齐策略算法的运行时间。具体实验结果如表 6。

表 4 各类语料中句对类型统计结果

语料类型	英文	中文	总句对数	1-0(0-1)	1-1	1-2(2-1)	2-2	其他
会议记录	1460	1497	1389	64	1110	191	7	17
新闻	462	411	405	74	229	78	2	22
法律	1399	1493	1461	95	1304	60	0	2
总计	3321	3401	3255	233	2643	329	9	41
所占比例(%)			100	7.2	81.2	10.1	0.3	1.2

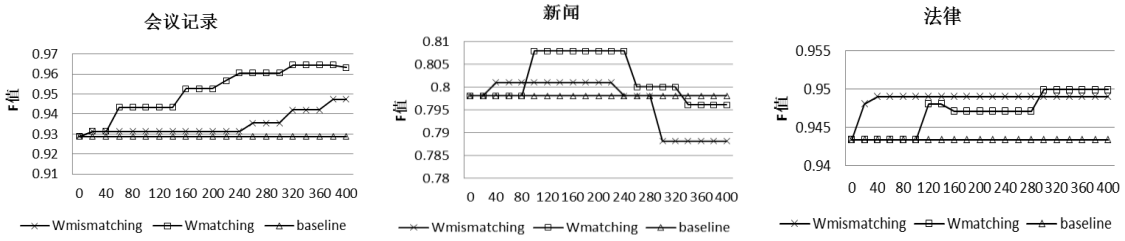

图 2 不同类语料中 $W_{mismatching}$ 与 $W_{matching}$ 对算法影响的对比结果

表 5 融合词对齐位置映射关系方法与 champollion 方法对比结果

语料类型	Champollion 句子对齐算法			融合词对齐位置映射关系的句子对齐方法		
	准确率	召回率	F 值	准确率	召回率	F 值
会议记录	0.9215	0.9362	0.9288	0.9505	0.9707	0.9605
法律	0.9892	0.9945	0.9918	0.9924	0.9967	0.9946
新闻	0.8037	0.8113	0.8075	0.8349	0.8271	0.8311

表 6 各类对齐算法运行时间对比结果

对齐方法	准确率	召回率	F 值	运行时间（s）
基于长度的对齐方法	0.8031	0.7729	0.7877	4
Champollion 对齐方法	0.9369	0.9496	0.9433	87
融合词对齐位置映射关系的方法	0.9526	0.9679	0.9602	526
结合锚点对齐策略后的方法	0.9526	0.9679	0.9602	46

对上述实验结果，本文进行了以下 6 点分析：

1) 根据表 2,本文使用的 match-byte 锚点选择方法，在保证准确率的前提下，大大提高了锚点识别的召回率，在三类测试语料上的实验结果相比 Li 的方法有显著提高。
2) 根据表 3，未对中文进行分词处理，有效的避免了由分词错误导致匹配失败的 21 现象，从而提高了锚点识别方法的召回率。按分词后单词的个数作为匹配文本长度，只是简单考虑了匹配成功词对的数量所占比例，且每个词的长度都是 1，即每个词的重要程度均相同。而采用字节方式，给予不同的词不同的权重，即长度较长的词语重要性更高，这也符合语言学中，较长词汇往往包含更多信息的现象。两者的结合大大提高了锚点的识别效率。
3) 根据图 2，相比单独引入未匹配词汇的词对齐位置映射关系，单独引入匹配词对的词对齐位置映射关系对算法效率的提升效果更佳，两者在法律和会议记录类等较规范文档中提升效果均较好且稳定，但在噪声较多的新闻类文档中稳定性较差。
4) 根据表 5，融合词对齐位置映射关系的方法,在多类语料中都有效地提高了句子对齐结果,平均 F 值提升了约 1.94%，其中在会议记录类语料中效果最为明显，F 值提高了约 3.18%。
5) 根据表 6，结合基于锚点对齐策略的方式大大提高了融合词对齐位置映射关系的句子对齐算法的速度，使算法面对

大规模语料对齐工作时变的更加实用。

6) 根据实验结果分析，$W_{matching}$ 和 $W_{mismatching}$ 对算法的影响较大，本文只是简单地选择在多类数据集上使算法效果都较优的参数值。使用更好的优化方式来优化两个参数会使算法效果更佳。

4 总结与未来工作

本文描述了一种融合词对齐位置映射关系的句子对齐方法。通过使用以字节个数作为词汇长度的计算方式，实现高效的锚点句对识别方法，在不失准确率的前提下，显著提高了锚点句对识别的召回率。再利用锚点句对将双语语料切分成多个较小的待对齐片段，在待对齐片段内使用融合词对齐位置映射关系的句子对齐方法进行对齐。实验结果表明，无论在规范语料还是噪声语料上，本文提出的句子对齐算法，都非常有效地提高了算法的对齐效果。同时词对齐位置映射关系这一新特征不仅只适用于中英句子对齐算法，同样还适用于其他语言的句子对齐算法和不同策略的句子对齐算法。但对于翻译较文学化的句对，如“With the passage of time, earthshaking changes have taken place. 斗转星移，沧桑巨变。”，此类人工都难以判断的句对，仍无法实现准确对齐。所以文本后续工作将加入人工指导来干预句子的对齐，探索用户干预点的选取技术，及人工指导后对整体对齐质量的影响。

参考文献

[1] Christopher D, Manning H S. Foundations of Statistical Natural Language Processing[M]. Electronics Industry Press, 2002.

[2] 王伟. 机器翻译中的对齐技术研究[D]. 北京邮电大学, 2002.

[3] Brown, P.E., Lai, J.C and Mercer, R.L.. Aligning Sentences in Parallel Corpora[C]. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics(ACL′91), 1991:169-176

[4] Gale, W. A., Chuich, K. W. (1991). A program for aligning sentences in bilingual corpora, in ACL ′91, Berkeley CaA, pp. 177-184

[5] 吕学强, 李清隐, 任飞亮,等. 基于统计的汉英句子对齐研究[C]// 第一届学生计算语言学研讨会论文集. 2002:990-992. DOI:10.3969/j.issn.1000-1220.2004.06.012.

[6] 张霞, 昝红英, 张恩展. 汉英句子对齐长度计算方法的研究[J]. 计算机工程与设计, 2009, 30(18):4356-4358.

[7] S.F. Chen. Aligning sentences in bilingual corpora using lexical information[C].Pro. of the 31th Annual Meeting of the ACL, 1993:9-16.

[8] Wu D. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria[J]. In ACL-94, 1994:80--7.

[9] 钱丽萍, 赵铁军, 杨沫昀,等. 基于译文的英汉双语句子自动对齐[J]. 小型微型计算机系统, 2001, 36(1):123-125. DOI:10.3321/j.issn:1002-8331.2000.12.023.

[10]Ma, Xiaoyi. Champollion: A robust parallel text sentence aligner. In Proceedings of LREC-2006: Fifth International Conference on Language resources and Evaluatioin, 2006, page 489-492.

[11]Trieu H L, Nguyen P T, Nguyen K A. Improving Moore′s Sentence Alignment Method Using Bilingual Word Clustering[M]// Knowledge and Systems Engineering. Springer International Publishing, 2014:149-160.

[12]Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora[M]// Machine Translation: From Research to Real Users. Springer Berlin Heidelberg, 2002:135-144.

[13]Peng Li, Maosong Sun, Ping Xue. 2010. Fast-Champollion: a fast and robust sentence alignment algorithm. In Proceedings of ACL 2010: Posters, pages 710-718.

[14]Maimaitimin S, Hou M. Chinese-Uyghur Sentences Alignment Using Multiple Clues[J]. Advanced Materials Research, 2014, 989-994:4990-4995.

[15]Ru C, Dan O5tef0Onescu, Dan T. Acquiscommunautaire sentence alignment using support vector machines[J]. In Proceedings of the Fifth Language Resources and Evaluation Conference (LREC, 2006, 46(4):2134--2137.

[16]Fattah M A, Ren F, Kuroiwa S. Probabilistic Neural Network Based English-Arabic Sentence Alignment[J]. Lecture Notes in Computer Science, 2006, 3878:97-100.

[17]Ghaly H. Canvas: A fast and accurate geometric sentence alignment system using lexical cues within complex misalignment settings[J]. Dissertations & Theses - Gradworks, 2014.

[18]熊伟, 陈蓉, 刘佳,等. 面向小词典的高效英汉双语语料对齐算法[J]. 计算机工程, 2007, 33(13):210-212.DOI:10.3969/j.issn.1000-3428.2007.13.072.