越汉双语句子自动对齐研究初步*

陈坚忠，李 鹏，孙茂松
智能技术与系统国家重点实验室
清华信息科学与技术国家实验室（筹）
清华大学 计算机系，北京 100084
E-mail: tktrungna@gmail.com; pengli09@gmail.com; sms@mail.tsinghua.edu.cn

摘 要： 句子级对齐双语语料是自然语言处理的重要资源之一，对于机器翻译、跨语言检索、双语词典编纂等研究有很大应用价值。关于自动句子对齐的研究主要针对于英语、法语、汉语等语言，据我们所知，尚未见到针对越南语-汉语的相关研究。本文考查了使用不同参数时，基于长度的句子对齐算法、Champollion 算法在越南语-汉语双语文本上的效果，并根据汉字与越南语音节间的独特对应关系对 Champollion 算法进行了改进，获得了更好的对齐效果。

关键词： 越汉句子自动对齐

Preliminary Study on Vietnamese-Chinese Bilingual Sentence Alignment

Kien Trung Tran, Li Peng, Sun Maosong
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084
E-mail: tktrungna@gmail.com; pengli09@gmail.com; sms@mail.tsinghua.edu.cn

Abstract: Sentence-level aligned parallel corpora are very important resources for a number of natural language processing tasks, including machine translation, cross language information retrieval and lexicography. In this paper, we investigate the performance of length-based sentence alignment algorithm and Champollion algorithm for Vietnamese-Chinese sentence alignment. And we propose a method to improving the Champollion algorithm by adopting the correspondence between Vietnamese syllables and Chinese characters. Preliminary experiments show the effectiveness of this method.

Keywords: Vietnamese-Chinese bilingual sentence alignment

1 引言

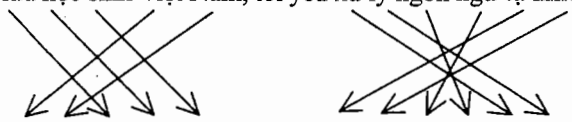
随着经济的发展，中越两国之间的交流、合作越来越多，越南语-汉语（简称越汉）双语相关信息处理需求也越来越强，如越汉机器翻译技术、跨语言检索技术等，相应的研究工作也蓬勃开展起来。越汉双语语料库，特别是句子级对齐的越汉双语语料，是这些研究工作的基础性资源，越汉双语语料库的构建技术具有重要的学术和商业价值。互联网上具有大量越汉双语网页，是越汉双语语料的重要来源，但这些网页多数都只是在篇章级对齐，手工找出这些篇章中句子间的对应关系（即“句子对齐”）费时费力，无法实用。因此利用计算机自动进行句子对齐对构建越汉双语语料库具有重要意义，但据我们所知，目前尚未见到越汉双语句子对齐的相关研究工作发表。在本论文中，我们考查了在其他语言对上常用的基于长度的句子对齐算法和 Champollion 算法在越汉语言对上的性能，并针对汉字与越南语音节间的独特对应关系对 Champollion 算法进行了改进，以解决汉语和越南语分词标准不一致带来的问题。

现代越南语文字采用拉丁字母，其基本组成单位是音节，音节间以空格进行分隔。在历史上的一段时期内，越南与中国之间有着比较深刻的接触，并以汉字作为自己的正式文字，因此汉语对越南语产生了很大影响。在现代越南语中，对于每个汉字，都有一个或多个音节与之对应，称

* 本文承国家自然科学基金（60873174）资助

为汉越音(Sino-Vietnamese)，这一现象是越南语与汉语间所特有的。例子 1 展示了一个越南语句子及其对应的汉语句子，以及该对句子中汉字与越南语音节之间的对应关系。

例子 1：Tôi là lưu học sinh Việt Nam, tôi yêu xử lý ngôn ngữ tự nhiên.



我 是 越 南 留 学 生，我 爱 自 然 语 言 处 理。

正如上文所述，越南语中的空格只作为音节间的分隔符，而不是词的分隔符，因此越南语像汉语一样存在分词问题。例子 2 中展示了例子 1 中句对理想的分词结果，表 1 列出了例子 2 中越南语词和汉语词间的对应关系。近十年来，学术界对越南语的分词问题进行了大量研究，目前已有一定的成果[4,5,6,8]。现代越南语词汇大致可分成三类：固有词（本身就有的词汇）、汉越词（自古汉语派生出的词汇）以及外来词（由古汉语以外的语言如英语、法语等传入的词汇）。其中，汉越词的意义和用法跟相应的现代汉语词并不完全一致，在不同类型文档中的比例也有所不同，如在科学、行政等领域的文档中比例较高，而在小说、新闻等领域的文档中比例则会低一些，但一般其比例不低于 60%[1]。

例子 2：【越南语】Tôi/ là/ lưu học sinh/ Việt Nam/,/ tôi/ yêu/ xử lý/ ngôn ngữ/ tự nhiên/.

【汉 语】我/是/越南/留学生/，/我/爱/自然/语言/处理/。

表 1 例 2 中越南语词与汉语词间的对应关系

越南语词	汉语词	越南语词	汉语词
lưu học sinh	留学生	ngôn ngữ	语言
Việt Nam	越南	tự nhiên	自然
xử lý	处理		

论文后续的内容安排如下：第二部分介绍了句子对齐有关的概念和相关工作，第三部分介绍了算法细节和我们所作的改进，第四部分给出实验结果和讨论，第五部分进行总结。

2 概念与相关工作

2.1 句子对齐相关概念

为了叙述方便，我们首先定义“句珠”和“互译单元”这两个概念。

句珠（bead）：一个句珠由一句或多句源文与一句或多句译文组成。在本文中，记 V、C 分别为越南语文档和汉语文档，我们用如下符号表示一个句珠：$A_i = (V_{A_i}, C_{A_i}) = (V_{a_{i-1}+1}, V_{a_{i-1}+2}, \cdots, V_{a_i}; C_{a_{i-1}+1}, C_{a_{i-1}+2}, \cdots, C_{a_i})$，其中 V_i、C_j 分别表示 V、C 的第 i 个和第 j 个句子。后文中我们用 m-n 表示一个句珠包含 m 个源文句子和 n 个译文句子。

互译单元：我们称一对互为翻译的字串为一个互译单元。互译单元的粒度比较灵活，可以是一对互译的汉字与越南语音节，如“lưu”与“留”，也可以是一对互译的词，如“lưu học sinh”与“留学生”，甚至可以是一对互译的句子片断，如“Tôi là lưu học sinh Việt Nam”与“我是越南留学生”。

2.2 基于长度方法

基于长度方法的出发点是：一般比较长的句子的译文也比较长，而比较短的句子的译文也比较短，从而可以利用源文与译文句子长度间的对应关系作为对齐的依据。

[1] http://zh.wikipedia.org/wiki/汉越词，（访问时间：2011 年 4 月 19 日）

在已有文献中，对于句子长度有两种度量方式：Brown 等人在文献[1]中认为翻译的基本单位应该是词，所以应以词数作为句子长度的度量单位；而 Gale 和 Church 在文献[2]中认为某些句子中所含的词的数目比较少，以词数作为句子长度的度量单位会使度量准确性变差，而以字节数作为度量单位的准确性会更高，所以应以字节数作为句子长度的度量单位。本文在实验部分对这两种度量方式都进行了考查。

基于长度方法优点在于存储开销小、运行速度快。对于一些比较相近的语言对，如英语和法语等，采用这种方法可得到比较好的结果[1,2]。但是，它只用了简单的长度信息而忽略了句子中的丰富词汇信息，所以对于语系上差别较大的语言对（如英汉）正确性有所下降[3]。

2.3 基于词汇信息方法

该类方法考虑了词汇信息在句子对齐中的作用，一般会比基于长度方法取得更好的效果。基于词汇信息方法又可细分为两大类：不使用词典的方法与使用词典的方法。不使用词典的方法适用于使用相似文字的语言对（如英语与法语），这些语言对中存在一定的同源词，可以利用启发式规则和字符串匹配来对这些同源词进行匹配，以帮助进行句子对齐。使用词典的方法借助双语词典实现语言间词汇的匹配，以帮助进行句子对齐，适用范围更广。越南语与汉语文字存在较大差别，无法通过简单的规则在词语间或音节与汉字间实现匹配，只适合使用基于词典的方法。

Champollion 算法[3]是 Xiaoyi Ma 在 2006 年提出的一种使用词典的句子对齐算法。它认为在确定句子是否互译的过程中，仅在少数文档中偶尔出现的互译词汇要比常常在很多文档中出现的互译词汇具有更高的置信度。基于这一观察，它参考信息检索中常用的 tf-idf 模型对互译词汇进行加权，取得了很好的效果。这一方法简单有效，因此本文中将其作为典型的基于词汇信息方法进行考查，并针对越南语与汉语的特点对其进行了改进。

基于词汇信息方法较基于长度方法存储开销要大，速度要慢，但具有更好的鲁棒性。另外，这一类方法的效果依赖于词典规模。采用的词典规模越大、质量越高，效果越好。所以准备比较好的词典是一个很重要的工作步骤。

3 越汉句子对齐实现

3.1 基于长度方法

基于长度方法为每一种可能的对齐结果赋予不同的概率，将概率最大的对齐作为最优对齐。概率模型的定义如下：

假设 V_{Ai}, C_{Ai} 互为翻译的概率只依赖于它们的长度属性，且句珠间是相互独立的，则 V、C 对齐的概率可表示为：$Prob(A|V,C) \approx \Pi_{i=1}^{I} prob(V_i \Leftrightarrow C_i | l_{V_{Ai}}, l_{C_{Ai}})$，根据条件概率公式：

$$Prob(V_i \Leftrightarrow C_i | l_{V_{Ai}}, l_{C_{Ai}}) = \frac{Prob(l_{V_{Ai}}, l_{C_{Ai}} | V_i \Leftrightarrow C_i) \times Prob(V_i \Leftrightarrow C_i)}{Prob(l_{V_{Ai}}, l_{C_{Ai}})}$$

$$\approx Prob(l_{V_{Ai}}, l_{C_{Ai}} | V_i \Leftrightarrow C_i) \times Prob(V_i \Leftrightarrow C_i)$$

（这里对于任意的 $V_{Ai}, C_{Ai}, Prob(l_{V_{Ai}}, l_{C_{Ai}})$ 可以认为是常数，所以可以忽略）

对于 0-1 或 1-0 型句珠，可以利用 V、C 中句子长度的分布来估计 $Prob(l_{V_{Ai}}, l_{C_{Ai}} | V_i \Leftrightarrow C_i)$。而对于其他类型的句珠，可进一步分解为：$Prob(l_{V_{Ai}}, l_{C_{Ai}} | V_i \Leftrightarrow C_i) = Prob(l_{C_{Ai}}) \cdot Prob(l_{V_{Ai}} | l_{C_{Ai}} | V_i \Leftrightarrow C_i) \approx Prob(l_{C_{Ai}}) \cdot \alpha \exp\left(-\frac{(r_i - \mu)^2}{2\sigma^2}\right)$ 其中 $r_i = \log(l_{V_{Ai}} / l_{C_{Ai}})$，$\alpha$ 为归一化因子，μ 与 σ^2 可从标注语

料库上统计得到。

本文中我们考虑了两种长度单位的定义，即音节/汉字和字节。计算方法为：对于越南语，以空格作为音节的分隔符，被空格分开的每一组越南语字母计一个音节，每个越南语的字母（如 a, b, ă, ắ, ...）计一个字节；对于汉语，每个标点也计作一个汉字，每个汉字计两个字节。

3.2 Champollion 算法

Champollion 算法定义了句珠的相似度，并将一个对齐中各句珠相似度的总和作为该对齐的评分，取评分最高的对齐作为最优对齐。

对于两段文本 V_{Ai}, C_{Ai}，设 $P=\{(v_1',c_1'),(v_2',c_2'),\cdots,(v_k',c_k')\}$ 为它们中的 k-互译单元集。借用信息检索中常用的 tf-idf 模型，对于每个互译单元对 (v_i',c_i')，定义 idtf (term frequency–inverse document frequency)、stf (segment-wide term frequency)如下：

$$idtf(v_i')=\frac{v_i'\text{在整个文档}V\text{中出现的频率}}{v_i'\text{在}V_{Ai}\text{中出现的频率}},\ stf(v_i',c_i')=\min\{stf(v_i'),stf(c_i')\},$$

其中 $stf(v_i'),stf(c_i')$ 分别为 v_i',c_i' 在 V,C 中出现的频率。V_{Ai}, C_{Ai} 的相似度评价函数 $sim(V_{Ai}, C_{Ai})$ 定义为：

$$sim(V_{Ai},C_{Ai})=\sum_{i=1}^{k} lg(stf(v_i',c_i')\times idtf(v_i'))\times alignment\_penalty\times length\_penalty,$$

其中 $alignment\_penalty=\begin{cases}1 & \text{对于1-1型名珠}\\ \text{大于0小于1的值} & \text{对于其他类型名珠}\end{cases}$，$length\_penalty$ 是关于 V, C 的长度的函数。

对于英语、法语等，句子中的最小单元是词，且可以简单的按空格来分词，然而对于汉语、越南语，最小单元分别是汉字和音节，且二者间除了词间的对应关系外还有音节与汉字间的对应关系，因此可以考虑将词或音节作为互译单元。相应的对每一种互译单元的定义，需要构造相应的双语词典（基于词的词典、基于音节/汉字的词典）。

虽然越南语、汉语的分词算法已经做得比较好，但两种语言的分词器采用的分词标准不同，分词结果也不同，这样当以词为互译单元时会出现找不到互译单元对的情况。例如在例子 2 中，有两个互译词对：（ngôn ngữ，语言）和（tự nhiên，自然），而 vnTokenizer[1]会把“ngôn ngữ tự nhiên”分为一个词，但 ICTCLAS[2]则把“自然语言”分成“自然”、“语言”两个词。这样通过词典就无法找出这两个词组的互译关系，导致例子 2 的两个句子相似度下降。如果定义互译单元为音节，通过音节词典可以找到两词组中音节互译关系为：（ngôn，言）、（ngữ，语）、（tự，自）、（nhiên，然），因此会把例子 2 的两个句子相似度提高。但越南语词汇中除了汉越词还有固有词、外来词，另外不少汉字对应的越南音在现代越南语中很少使用，例如在例子 3 中，汉字 “胶”对应的越南音是“giao”，而句子中使用的是“cao su”这个词，需要使用基于词汇的词典来对这个互译对进行验证。

例子 3：　【越南语】Ông ta/ đi/ dép/ cao su/ lên lớp/./

　　【汉　语】他/穿/了/胶鞋/上/课/。/

为了解决这一问题，我们提出一个改进方法。借用前向最大匹配分词方法的思想，我们利用词典来“分词”，找出互译单元。设 $V=\{v_1, v_2, \cdots, v_{m-1}, v_m\}$、$C=\{c_1, c_2, \cdots, c_{n-1}, c_n\}$ 分别为越南语文本与汉语文本，v_i, c_j 为相应文本中的音节或汉字。找出对应互译的单元算法如算法 1 所示。后文中，我们将使用“改进音节”指代用此算法找出的互译对。

[1] http://vlsp.vietlp.org:8080/demo/dl/VnTokenizer_VLSP_SP82_20100804.tgz

[2] http://ictclas.org/index.html

（1）考虑源文中的音节 (v_i)
（2）if 存在 $(c_k, c_{k+1}, \cdots, c_{k+l}) \subseteq C$ 并且 $(v_i) \leftrightarrow (c_k, c_{k+1}, \cdots, c_{k+l})$ 在词典中出现
then 当为一个互译对，且接着考虑源文中音节 (v_{i+1})，跳到(1)
（3）else 考虑源文中的两个音节 (v_i, v_{i+1})
（4）if 存在 $(c_{k'}, c_{k'+1}, \cdots, c_{k'+l'}) \subseteq C$ 并且 $(v_i, v_{i+1}) \leftrightarrow (c_{k'}, c_{k'+1}, \cdots, c_{k'+l'})$ 在词典中出现
then 当为一个互译对，且接着考虑源文中音节 (v_{i+1})，跳到(1)
（5）else 继续考虑源文中的三音节 (v_i, v_{i+1}, v_{i+2})
一直考虑到 $(v_i, v_{i+1}, \cdots v_{i+max-1})$，其中 max 为词典中源文词最大的长度。
（实验中使用的词典中词的长度一般小于等于 2，所以设 max = 2）

算法 1　基于词汇改进句子对齐方法

4　实验结果与讨论

4.1　测试语料与评价方法

我们从网上收集了 31 篇越汉双语文章作为测试语料，这些文章涵盖政府文档、短篇小说、新闻、专业文献等多种体裁，共包含 1540 个越南语句子，1514 个汉语句子。我们对这些文章进行手工对齐，得到 1474 个句珠，作为标准答案。标准答案中各种类型句珠的比例见表 2。

表 2　数据集中句珠类型的统计

类型	数量	比例(%)
1-1	1349	91.52
1-2 或 2-1	99	6.72
0-1 或 1-0	15	1.02
2-2	11	0.75
总计	1474	100.00

表 3　基于长度两种方法的参数

基于长度方法	μ	σ^2
字节为长度单位	0.5337	0.0509
音节/汉字为长度单位	-0.1031	0.0511

基于词汇信息方法很重要的资源是双语词典。词典的质量会影响到对齐结果。在下文的实验中我们考查了以音节/汉字和词分别作为互译单元时 Champollion 算法的性能，并相应构造了两部词典：第一部是基于音节/汉字的词典[1]，包括 15741 个词条；第二部是基于词的词典[2]，包括 92496 个词条。

我们使用 $\text{Prencision} = \frac{|\text{GB} \cap \text{PB}|}{|\text{PB}|}$、$\text{Recall} = \frac{|\text{GB} \cap \text{PB}|}{|\text{GB}|}$、$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{|\text{Precision+Recall}|}$ 和运行时间衡量句子对齐算法的性能，其中 GB, PB 分别是手工标注和经过自动对齐过程得到的句珠集。

4.2　实验参数和条件

对于基于长度方法，我们考查了以音节/汉字、字节分别作为句子长度单位情况下算法的性能。对应不同的长度定义，构造的概率模型的参数不同。表 3 为 μ, σ^2 参数的值，其中 μ 通过计算标准答案各句珠中越汉句子长度比值的平均值的对数得到，σ^2 通过计算标准答案各句珠中越汉句子长

[1] 数据来自 Hán Việt Tự Điển 电子版 www.hanviet.org
[2] 数据来自 MTD Vietnamese – Chinese 词典 www.lacviet.org

度比值的平均值的对数的方差得到。

对于 Champollion 算法，我们分别考查了以词、音节、改进音节作为互译单元时算法的性能。互译单元为音节时使用基于音节的词典，其他情况使用基于词的词典。实验中用到的 *alignment_penalty*, *length_penalty* 借用 Champollion 算法的 Perl 实现[1]中定义的形式。

在以下实验中，我们使用 vnTokenizer4.1.1c 作为越南语分词工具，文献[6]中报告的准确率达到96%；使用 ICTCLAS 作为汉语分词工具，ICTCLAS 的主页提到准确率达到 98.45%。

类似于英语、法语等，越南语中的句号有许多歧义，需要对句子边界进行辨识。在文献[5]中，作者提出了基于最大熵原理的越南语句子边界识别算法，得到了较好的结果（论文中报告的召回率为 95%），并提供了辨识工具 vnSentDetector（vnTokenizer 包的插件）。在以下实验中，我们直接使用这一工具划分越南语句子，而使用句号划分汉语句子。

4.3 不同算法在测试语料上的性能

各算法在测试语料上的结果如表 4 所示。从实验结果中可以看到，基于长度方法在速度上占优势，也获得了很好的 Precision、Recall 和 F-measure。以词作为互译单元时，Champollion 算法的运行时间比较长，对齐结果也较差。原因是此方法需要经过耗时的越南语、汉语分词过程，使得运行时间变长。另一方面两种语言的分词工具的分词标准不同，与词典中的词条的划分标准也并不完全一致，导致部分出现在双语句子中的互译词因切分不一致而无法在词典中检索到，从而使对齐结果变差。以音节作为互译单元时，不需要经过分词过程，同时基于音节的词典更小，所以运行比较快。以改进音节作为互译单元时，在较短的运行时间内，获得了最高的 Precision、Recall 和 F-measure，在对齐效果与运行时间间获得了平衡。

表 4 各算法在测试语料上的性能

方法	Precision	Recall	F-measure	运行时间(s)
基于长度-字节为长度单位	0.9546	0.9559	0.9552	2.329
基于长度-音节/汉字为长度单位	0.9489	0.9444	0.9466	3.841
Champollion-互译单元为词	0.8603	0.8901	0.8750	41.592
Champollion-互译单元为音节	0.9443	0.9552	0.9497	7.941
Champollion-互译单元为改进音节	0.9769	0.9749	0.9759	12.836

4.4 算法鲁棒性

在文献[2,3]中报告了相应的句子对齐算法在 0-1 或 1-0 类型（非直译或省略句子）的句珠上的性能较其他类型句珠要差，图 1 展示了五种算法在测试语料的不同类型句珠上的 F-measure。可以看到，对于越汉语言对，同样的问题依然存在。

互联网上的双语文本中经常会出现非直译或省略句子的情况，因此算法在 0-1 或 1-0 类型句珠上的性能对于算法能否实用具有很大影响。为了考查各算法的鲁棒性，我们设计了如下实验：在测试语料中随机插入一些无关

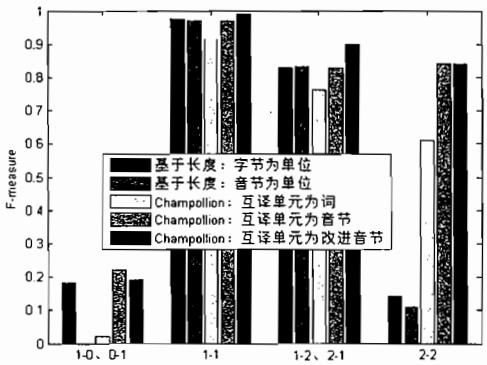


图 1 各种句珠对应的 F-measure 比较

[1] http://champollion.sourceforge.net/

句子，以提高标准答案中 0-1 和 1-0 型句珠的比例，通过观察不同算法在这些语料上的性能来考察算法的鲁棒性。图 2 展示了当 0-1 和 1-0 型句珠的比例分别为 1%、5%、10%、15%、20%、25%、30%时，不同算法在这些语料上得到的 F-measure。可以看到，当此比例升高时，各算法的性能都有下降，而基于长度方法的性能下降得更快。这说明基于长度方法虽然简单、速度快，但不适用于噪声较多的双语语料。以词、音节或改进音节作为互译单元的 Champollion 算法具有更好的鲁棒性，适用情形更广。

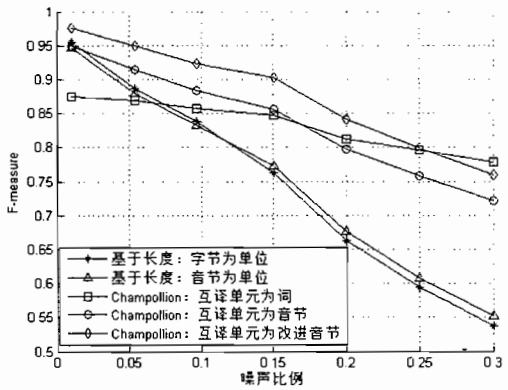


图 2　各种方法的鲁棒性

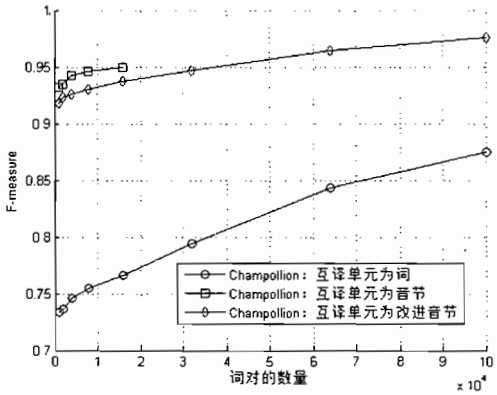


图 3　词典规模对基于词汇三种方法的影响
（基于词汇方法定义互译单元为音节的词典规模最多是 15741 词条）

4.5　词典规模对 Champollion 算法的影响

词典对于基于词汇方法有很大的作用。当观察两句子时，如果从词典中找到更多互译单元对，此两句子是互译的可能性更大。为了观察词典规模对于基于词汇方法的影响，我们仿照文献[3]中的方法构造了不同规模的词典，即先将词条按中文部分在人民日报语料库[1]中出现的频率从高到低排序，再取前 K 个词条，构成词典。图 3 是实验结果，可见词典的规模越大，对齐结果越好。通过统计发现，对于含有 15741 个词条的词典，只有 20%的词条出现在标注语料库的句对中，其他的词条或者本身就很少使用，或者实际使用的译文与词典给出的译文不一致。所以为了提高对齐效果，构造一部高质量词典是很重要的。

综合以上实验，我们可以看到，以改进音节作为互译单元的 Champollion 算法在对齐效果、运行速度、鲁棒性三者间取得了较好的平衡，是本文所考查的五种算法中最适于应用于互联网环境的算法。

5　结语

本文介绍并考查了基于长度方法和 Champollion 算法在不同条件下，在越南语-汉语语言对上的效果，并针对越南语与汉语间特有的音节与汉字对应关系，提出了以改进音节作为互译单元的 Champollion 算法，获得了更好的效果。

如 4.5 节所述，双语词典的质量对于 Champollion 算法的性能具有很大影响，如何自动构建双语词典、提高词典质量是我们未来的研究工作之一。此外以改进音节作为互译单元的 Champollion

[1] http://www.icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp

算法的速度仍然较慢，不适合处理大规模双语文本，如何提高该方法的运行速度也是我们未来的研究工作之一。

参 考 文 献

[1] Brown, P. F. and Lai, J. C. and Mercer, R. L. 1991. Aligning Sentences in Parallel Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, p169-176.

[2] Gale, W. A. and Church, K. W. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistic, vol. 19, no. 1 March 1993*, p75-102.

[3] Ma, X. 2006. Champollion: A Robust Parallel Text Sentence Aligner. *Proceedings of Fifth International Conference on Language Resources and Evaluation*, p489-492.

[4] Ha, L. A. 2003. A Method for Word Segmentation in Vietnamese. *Proceedings of the International Conference on Corpus Linguistics, Lancaster, UK*, p282-287.

[5] Phuong, L. H. and Vinh, H. T. 2008. Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts. *Proceedings of the IEEE International Conference on Research, Innovation and Vision for the Future, Vietnam.*

[6] Phuong, L. H. and Huyen, N. T. M. and Azim, R. and Vinh, H. T. 2008. A Hybrid Approach to Word Segmentation of Vietnamese Texts. *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain. Springer LNCS 5196, 2008*, p240-249.

[7] Dien, D. and Kiem, H. 2003. POS–Tagger for English – Vietnamese Bilingual Corpus. *Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics*, p88-95.