汉藏短语抽取

诺明花[1,2]，张立强[1]，刘汇丹[1,2]，吴健[1]，丁治明[1]

1. 中国科学院软件研究所，北京 100190

2. 中国科学院研究生院，北京 100190

E-mail:{minghua,zlq,huidan,wujian,zhiming}@iscas.ac.cn

摘 要： 本文将从汉藏法律法规和公文领域平行语料中提取双语短语对。考虑现阶段藏文资源不足，提出两步汉藏短语抽取方法。第一步是提取汉语语块，这部分工作不是本文工作重点。第二步是获取待翻译汉语短语的译文，该模块提出藏文词序列相交算法抽取藏文短语。该算法可以很好的抽取 1-1 和 1-n 连续和非连续藏文短语。

关键词： 汉藏短语抽取；藏文信息处理；中文信息处理

Chinese Tibetan Phrase Extraction

Minghua Nuo[1,2], Liqiang Zhang[1], Huidan Liu[1,2], Jian Wu[1], Zhiming Ding[1]

1. Institute of Software, Chinese Academy of Sciences, Beijing, 100190

2. Graduate University of the Chinese Academy of Sciences, Beijing, 100190

E-mail:{minghua,zlq,huidan,wujian,zhiming}@iscas.ac.cn

Abstract: This paper describes a method to extract phrase pairs from domain-specific Chinese-Tibetan bilingual corpus in laws and regulations and official documents. So far, widely used phrase extraction method heavily depends on the result of word alignment or additional resources like part-of-speech or syntactic analysis and so forth. Taking account of inadequate resources in Tibetan at present, this paper proposes two-phase Chinese-Tibetan phrase pairs extraction method. First step is extraction of Chinese phrase (multi-word chunk) using Nagao's Algorithm and Substring Reduction Algorithm. The second step is extraction of candidate Tibetan translation for translation-ready Chinese phrase. This paper proposes Tibetan words sequence intersection algorithm (TIA) to extract Tibetan phrase. TIA works well on not only continuous or discontinuous Tibetan phrase but also 1-1 translation or 1-n translation.

Keywords: Chinese Tibetan Phrase Extraction; Tibetan information processing; Chinese information processing

1 引言

计算机辅助翻译的重要思想(包括基于翻译记忆技术和基于实例模式的翻译技术)是在翻译记忆库（双语对齐库）和实例模式库中搜索相同或相似的句子或短语，给出参考译文。基于实例的翻译，在翻译过程中使用一个搜索和匹配算法在平行语料库中寻找最优匹配的翻译实例，根据该实例的译文构造当前所翻译单元的译文。短语译文获取作为构造翻泽单元中未匹配部分的主要方法，是EBMT中不可缺少的核心环节之一。本文工作要为特定领域汉藏多策略机器辅助翻译系统构建短语对齐库。

基于短语的统计机器翻译的优势在于短语能够抓住局部上下文的依赖关系。因此，近几年来短语对抽取一直是统计机器翻译的重要模型之一。迄今为止，已经出现了多种短语抽取方法[1-8]。其中，有些计算复杂度太高，代价很高；有些模型依赖于词对齐的结果，有些依赖于句法分析结果，对资源的要求很高。汉藏短语抽取工作是史无前例的，本文要从汉藏双语平行语料中抽取互

译短语对。考虑目前还没有词性、句法层面上加工过的语料，汉藏短语获取方法一定摆脱对词对齐、句法分析等资源的依赖。本文获取的短语是广义上的，它是由若干个单词组成的语块。本文的思路是先从句对齐双语语料中获取有效汉语语块，对包含待译语块的句对求交集，得到候选译文，最后经过后处理得到汉藏互译语块。

2 翻译基本模型

Wang[9] 提出了一种基于序列相交的短语译文获取方法，该方法将句子视为词的序列，利用对中日句对齐语料库中包含待译短语的所有源语句子对应的目标语句子进行序列相交的方式，在不需要词对齐、句法分析及词典等资源的情况下，通过充分挖掘句对齐双语语料库的信息，获得高质量的短语译文。该方法不依赖于额外资源信息的特点正符合目前藏文多样化资源不充足，可以借鉴到特定领域多策略汉藏辅助翻译系统（简称MSCT_CAT）的短语库构建模型中。本节介绍藏文词序列相交算法译文获取模型。

2.1 基本模型

Wang[9] 提出的基于序列相交的短语对抽取方法，引入支持度的概念，仅仅考虑支持度排序结果中前两个译文作为备选，这种备选中往往会遗漏 1-n 的互译对藏文候选。然而，汉藏语料中 1-n 互译对是不可忽略不计的。例如：“第二款”在语料中有两种译法“ནང་གསལ་དཀྱུ་ཚན་” 和 “དཀྱུ་ཚན་”，其中译文二是译文一的子串，求交集过程中译文一很容易被忽略，或被译文二取而代之。因为两种译法的分布不均匀，完全用支持度远远不够。本文要从 Wang 提出的词序列相交的思想出发，针对藏文特点进行扩展。

藏文词序列相交算法（简称 TIA）使用的语料库为汉藏句对齐双语语料库 *SABC*，其中包含若干个汉藏对齐的句对。汉语句子是没有像英文那样自然形成的分词标记。作为一种拼音文字，藏文中各音节之间由音节点分隔，但是词与词之间没有分隔标记[10]，很难区分词的边界。为了词序列相交，本文分别使用斯坦福的中文分词开源项目和课题组开发的藏文分词模块对汉藏单语语料进行分词，为 TIA 做准备。TIA 的核心是通过汉藏词序列相交模型，获取 1-n 的汉藏互译短语对。

基本模型中，句子和短语均以词序列的形式表示。句子和短语的序列表示以及句子的序列相交定义沿用[9]的公式表示，表1给出汉藏双语句对词序列相交的示例。

表1 汉藏双语句对词序列相交示例表

SP_1	CS_1: 本法, 所, 称, 的, 国家, 通用, 语言, 文字, 是, 普通话, 和, 规范, 汉字。 TS_2: བཅའཁྲིམས་, འདི་, ར་, རྒྱལ་ཁབ་, ཀྱི་, སྐད་ཡིག་, རྣམ་, པ་, དེ་, སྐད་གོ་, དང་, མཐུན་རྐྱེན་, ཀྱི་, རྒྱུན་གྱི་, ཡིག་, །
SP_2	CS_1: 国家, 推广, 普通话, , , 推行, 规范, 汉字。 TS_2: ལ་ཚད་, ཀྱིས་, སྐྱོང་, ཞུགས་ཀདུལ་, དུ, གཏོར་, རྒྱུ, ར་, །, མཐུན་, ཀྱི, རྒྱུན་གྱི་, ཞུགས་སྐྱོང་, དུ, གཏོར་, རྒྱུ །
$CS_1 \cap CS_2$	{(普通话), (规范, 汉字)}
$TS_1 \cap TS_2$	{(སྐྱོང་), (མཐུན་, ཀྱི, རྒྱུན་གྱི་)}

2.2 基于序列相交的短语译文获取模型

先来分析一下表 1 中互译句对结构。例句表明，如果两个中文句子有公共子串，用 *CCS* 表示，它们对应的译文句子也有公共子串，用 *TCS* 表示，并且 *CSS* 与 *TCS* 在意义上是互译的。如果 *CSS* 刚好是待翻译汉语语块，通过计算藏文句子的公共子串的方法可以获取汉语语块的译文。

至于两组词串的对应关系，借用汉藏词典[11]可以确定两组汉藏互译短语对（“普通话”，ཕྱི་སྐད།）和

（“规范汉字”，ཚད་ལྡན་ཡིག་གཟུགས།）。

从以上分析可以得出，两个句对SP_r与SP_t相交结果表示如下：

$$SP_r \cap SP_t = (\{Q_1,Q_2,\cdots,Q_k\},\{T_1,T_2,\cdots,T_g\}) \quad (1)$$

$Q=\{Q_1,Q_2,\cdots,Q_k\}$ 为句对SP_r和SP_t中汉语句子CS_r和CS_t的交集（汉语短语集合），其中包含Q_i（$1\leqslant i \leqslant k$）待翻译的中文短语，$T=\{T_1,T_2,\cdots,T_g\}$ 为SP_r和SP_t中藏文句子TS_r 和 TS_t 的交集。T 中肯定包含Q_i的翻译译文，可以通过汉藏词典确定 (Q_i,T_j) 汉藏互译对。

待翻译中文短语由多个汉语单词构成，表示为如下公式（2）：

$$Q_i=<Q_{i+1},Q_{i+2},\cdots,Q_{i+l}> \quad (2)$$

假设Q_i中任意单词$Q_{i+\theta}$($1\leqslant \theta \leqslant l$)在词典中查到一个以上译文，所有译文保存到一个链结构L中，一定会存在某个$T_{j+\omega}$能够满足$T_{j+\omega}\cap L \neq \Phi$的条件。这些$T_{j+\omega}$($1\leqslant \omega \leqslant g$)最终构成$Q_i$的译文$T_j$。$T_j$可以是连续的，也可以是非连续的。

从公式（1）得知，句对的序列相交由若干个藏文公共子串*CS* 组成。其中为每个*CS* 构造一个树结构*T* 的话，句对的序列相交可以组成一个森林。*T* 由两种结点组成。

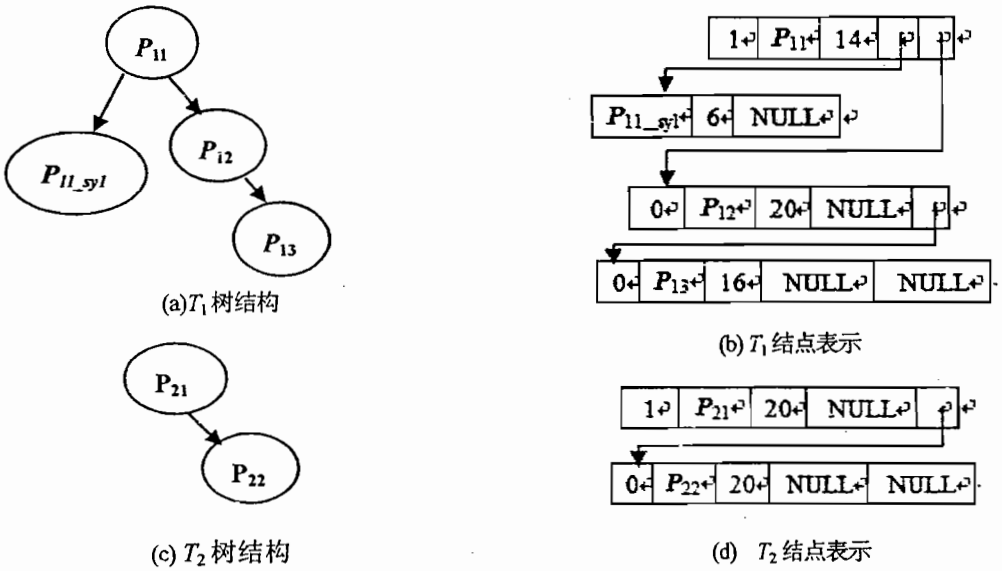
存储藏文句子取交后的某个*CS* 的单词结点用*ITN* 表示，在其左孩子结点中记录与*ITN* 链接的同义词，用*SYN*表示；右孩子结点中记录*CS*中相邻的单词。因此，某个*T* 的根结点是tag 域为1的*ITN* 结点，*T* 的叶子是有孩子为空的*ITN* 结点。*CS* 中某个单词对应结点的左子树由其所有同义词构成，形成一个同义词列表*SL*。

假设，在*SABC*中有20个中文句子包含待翻译语块Q ，其对应藏文句子取交后获取两个公共子串P_1 和 P_2。P_1 和 P_2的树结构分别用T_1 和 T_2表示，如图1。

确认(Q_i,T_j)的过程是对由 T_1 和 T_2 组成的森林的搜索过程。在图 1 中，P_{11_syl} 是 P_{11} 的同义词，同时 P_{11_syl} 和 P_{11} 出现频次的和等于 20，因此同义词$[P_{11},P_{11_syl}]$ 被接受。P_{12},P_{21} 和 P_{22} 出现频次均为 20，它们被接受为 T_j 的一部分。P_{13} 出现频次小于 20，从而被丢弃。Q_i 的最终翻译结果 T_j 是一个集合 P = { $[P_{11},P_{11\_syl}]P_{12},P_{21}P_{22}$}。

3 汉藏短语抽取

为了不依赖于额外资源，本文提出两步抽取汉藏短语方法。汉藏短语对抽取流程如图2所示。汉语和藏文语块抽取先后分两步来进行。在面向中文信息处理的研究工作中，吕学强和张乐[12]利用Nagao的N-gram 统计算法，在大规模汉语语料中进行抽取语块的实验，他们在论文中还提出一个删除同频子串的算法（SSR），提高了语块抽取的准确率。SSR可靠并复杂度不高，大语料处理中很使用。从汉语语块抽取的实际需求出发，本文在Nagao的串频统计方法的基础上开展



(a)T_1 树结构　(b) T_1 结点表示

(c) T_2 树结构　(d) T_2 结点表示

图 1 译文确认过程

基于词语的中文语块抽取并删除同频词串。提取的中文语块是连续的。具体串频统计和删除同频词串不是本文的重点，不再详细讲述，可以参考[12] [13]。

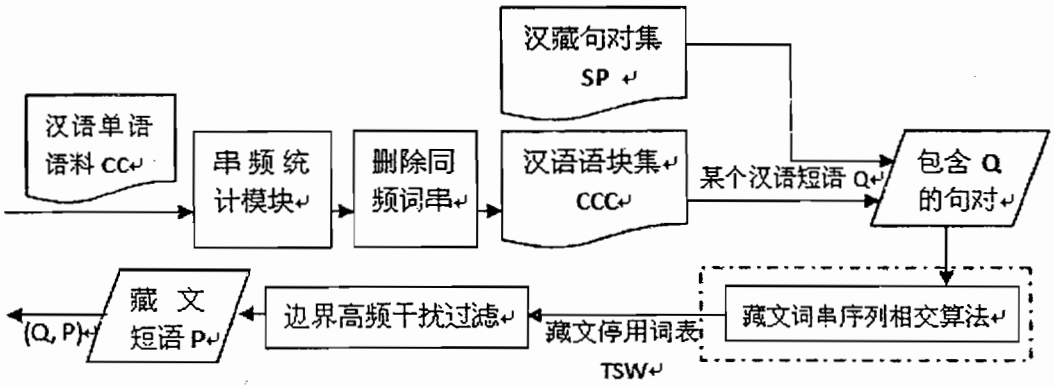


图 2 汉藏短语对抽取流程

除了藏文语块抽取（虚线内部）外汉藏短语对抽取需要做的工作有以下几点。虚线内算法在3.2节详细说明。首先，用脚本程序将*SABC* 分为汉藏各自的单语语料，汉语语料和藏文语料分别标记为*CC* 和*TC*。其次，用Nagao的算法计算出*CC* 中所有2-gram 到5-gram 语块做为候选汉语连续语块。根据[12]中算法，通过子串归并删除同一频度的子串。最后对这些候选汉语语块进行过滤和排序后将汉语语块集*CPS* 保存到文本文档中。为边界高频干扰过滤模块，构建*TC* 中的藏文停用词表*TSW*。

3.1 藏文短语类型

在藏文实际文本中，短语大体表现有两个层面的特性。首先有连续和非连续两种形式，称为短语连续性。例如：“རྒྱལ་ཁབ་ཀྱི་ཚད་གཞི་”是连续短语，“国家标准”的意思。由于翻译过程是根据上下文意译，容易形成非连续藏文短语。例如：“ཨོན་རྡལ་ཀྱི་སྒུལ་ཀ་བཟང་པོའི་”和“ཨོན་རྡལ་ཀྱི་སྒུལ་ཀ་བཟང་པོ་”，红色标出的是“产品质量认证”的译文，第一种译文中两个名词“ཨོན་རྡལ་ཀྱི་སྒུལ་ཀ་”和“བཟང་པོའི་”之间加了格助词[10]“ར”；第二层面是互译对之间 1-1 和 1-n 两种关系形式，称为短语对应关系。翻译机构的地域性引起同一个中文不同意法，类似一词多义，例如：“款”这个单词有的地区译成“རྡོག་ཞལ་དཔོན་ཚན་”，有的地方译成“དཔོན་ཚན་”，从而语料中两种甚至多种译法并存。短语对抽取过程中，一个中文多种译法的现象是不可忽略的。

3.2 藏文词串序列相交算法

该方法不依赖与额外资源的前提下，对句对齐双语语料库中包含待翻译汉语语块 Q 的句对 *TSS* 求交集，并通过后处理得到汉语语块的译文，从而构建汉藏短语库。为了提高准确率该方法用到汉藏词典[11]。TIA 重点解决 1-n 的短语对。

TIA 算法的核心由两步组成。第一步使用第二节介绍的序列相交翻译模型，对藏文句子集中任意两句取交来为已知的 Q 构建公共子串森林 F。由公式（2），Q 由若干个词 $Q_i(1 \leqslant i \leqslant l)$ 组成。取交过程中任意 Q_i 的译文均被保存并生成公共子串树 T 或森林 F。并不是 T 或 F 中所有结点构成 Q 的译文 P ，结点满足以下两个条件才是组成 P 的候选。

1）译文中一定包含任意 $Q_i(1 \leqslant i \leqslant l)$ 的译文；

2）所有候选译文的支持度和等于 S_n。

第二步遍历 F，筛选出满足以上条件的候选单词并确认 Q 的译文 P。P 是 *CS* 的集合，P 的生成过程描述如下。伪代码中用 A，B，C，D 分别将 P 标记为 1-1，1-n，连续或非连续等不同短语类型。

```
1.  int t_n =0, i=0.
2.  for each T in F
3.          for each tw in T
4.                  if freq equals S_n
5.                          add tw to{ Pi} and i++.
6.                  else if sum of freq in SL equals S_n
7.                          add sy_tw of SL to {Pi} and i++.
8.                  else
9.                          discard T.
10.         end for
11.       t_n ++
12. end for
13. if t_n ==1
14.         (Q,P) are marked as A
15. else if t_n >1
16.         (Q,P) are marked as B
```

17. If $i==1$
18. P is marked as C
19. else if $i>1$ then
20. P is marked as D

从公共子串树和森林结构和以上伪代码可以得出，TIA 抽取的短语既能满足藏文短语的连续性，又能满足短语对应关系。因此可以达到 MSCT_CAT 的短语库构建要求。

4 实验

实验数据是汉藏法律法规和公文报告等特定领域语料。收集到的原始语料通过篇章对齐和句子对齐后，再抽取单语语料，最终形成短语对抽取模块可以处理的五份汉藏单语语料。语料 1 和语料 4 中低频短语对（在语料中出现次数很少）较频繁，语料 5 在五组语料中句对数最多。

4.1 藏文短语连续性验证

在实验中，对五组语料用 TIA 进行短语抽取之后，采用计算机辅助人工的方法判断互译对正确与否，表 2 列出 TIA 抽取的连续短语和非连续短语统计结果。表 2 中 D 表示 Discontinuous，C 表示 Continuous。

表2 TIA抽取结果的连续性统计表

语料	C	准确率	D	准确率
语料 1	376	0.758	18	0.833
语料 2	101	0.832	3	1.0
语料 3	199	0.809	10	0.8
语料 4	353	0.762	24	0.833
语料 5	1132	0.840	57	0.859

4.2 藏文短语对应关系验证

为了证明 TIA 抽取 1-n 短语对的有效性，对 TIA 抽取到的结果分析其汉藏对应关系。表 3 显示对应关系分布情况。该方法获得的短语译文准确率均值达到 80%。

表3 TIA抽取结果的对应关系统计表

语料	1-1	准确率	1-n	准确率
语料 1	381	0.756	13	0.846
语料 2	104	0.894	0	
语料 3	201	0.806	8	0.75
语料 4	366	0.762	11	0.818
语料 5	1166	0.839	23	0.826

很显然，语料中非连续短语对和 1-n 的短语对并不可以忽略不计的。TIA 可以抽取连续和非连续的短语对。同时能够有效抽取 1-1 和 1-n 汉藏短语对。从实验结果分析，由于数据稀疏问题，语料 1 和语料 4 两组准确率在同组试验中低于其它语料。低频短语在序列相交过程中很容易带着额外的与译文无关内容，这些干扰信息导致这两组的准确率降低。设定频率限度可以提高准确率，损失召回率。在每组实验结果中，语料 5 的准确率最佳，这表明通常语料变大可以提高覆盖率，较高的覆盖率能提高准确率。

5 结束语

目前汉藏语料资源不足的前提下，本文提出了两步抽取汉藏语块的方法。第一步利用 Nagao 的 N-gram 统计算法和吕学强的 SRR 抽取有效汉语语块。第二步计算包含待译汉语语块的汉藏句对公共子串的思想出发，尝试藏文词串序列相交算法获取译文。其结果能满足多策略汉藏辅助翻译系统的记忆库建设需求。然而，目前收集的汉藏并行语料中存在数据稀疏问题。N-gram、SSR 以及 TIA 都是依赖于统计的，对于数据稀疏问题无济于事。改进语料覆盖率有利于扩建汉藏短语库。另外，法律法规语料中表示章节和时间的短语占一定比例，通过模板可以很好的提高准确率。因此短语模板的研究对 MSCT_CAT 很有价值。

参考文献

[1] Daniel Marcu, William Wong. A Phrase-based, Joint Probability Module for Statistical Machine Translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, PA, USA. July 2002.

[2] Dekai wu . Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics 1997. 23(3): 377—404.

[3] Ying Zhang, Stephan Vogel, Alex Waibel. Integrated phrase segmentation and alignment algorithm for statistical machine translation. In: Proceeding of International Conference on Natural Language Processing and Knowledge Engineering. Beijing, 2003.

[4] Ying Zhang, Stephan Vogel. Competitive Grouping in Integrated Phrase Segmentation and Alignment Model. In: Proceeding of ACL Workshop On Building and Using Parallel Texts. Ann Arbor, 2005, 159—162.

[5] H Kaji, Y Kida, Y Morimoto. Learning Translation Templates from Bilingual Texts. In: Proceedings of the 14th International Conference on Computational Linguistics. Nantes France, 1992, 672—678.

[6] Franz Josef Och , Hermann Ney . The alignment template approach to statistical machine translation. Computational Linguistics, 2004, 30(40)l 417-449.

[7] David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics. Ann ArTr, 2005.

[8] 何彦青，周玉，宗成庆，王霞. 基于“松弛尺度”的短语翻译对抽取方法. 中文信息学报，2007，21(5)：91-95.

[9] 王辰，宋国龙，吴宏林，张俐，刘绍明. 基于序列相交的短语译文获取. 中文信息学报，2009，25(1)：39-43.

[10] 周季文，傅同和. 藏汉互译教程. 北京，民族出版社，1999.

[11] 张怡苏. 藏汉大辞典. 北京，民族出版社，1993.

[12] Xueqiang Lv, Le Zhang, and Junfeng Hu. Statistical Substring Reduction in Linear Time. In: Proceedings of IJCNLP-2004, 2004.

[13] Nagao,Makoto, Shinsuke Mori. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In COLING-94, 1994.