基于关联度的汉藏多词单元等价对抽取方法*

诺明花[1,2]，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 100049
E-mail: {minghua, huidan, wujian, zhiming}@iscas.ac.cn

摘 要：针对为汉藏辅助翻译系统建立汉藏多词单元翻译词典这一任务，本文提出了 CMWEPM 模型。该模型首先依据关联度和结合度来确定汉语语料中多词单元的边界，然后根据词对齐信息分别抽取严格和约束汉藏多词单元等价对。CMWEPM 模型根据不同长度和频次对多词单元进行分类，并为不同类型设定不同阈值，最终提高了汉藏多词单元等价对的召回率，从而能够间接地提高汉藏辅助翻译系统的翻译质量。

关键词：藏文信息处理；多词单元；关联度

Collocation Based Chinese-Tibetan Multi-word Equivalent Pair Extraction Method

Nuo Minghua[1,2], Liu Huidan[1,2], Wu Jian[1], Ding Zhiming[1]
[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190
[2] Graduate University of the Chinese Academy of Sciences, Beijing 100049
E-mail: {minghua, huidan, wujian, zhiming}@iscas.ac.cn

Abstract: This paper aims to construct Chinese-Tibetan multi-word equivalence dictionary for machine-aided translation system. It proposes CMWEPM model which extract multi-word equivalences in two phases. First, CMWEPM defines the boundary of Chinese multi-word units by collocation and binding degree. Then it extracts strict or constrained multi-word equivalences based on word alignments respectively. CMWEPM model classifies multi-word units and set different thresholds for different types. This strategy improves the translation quality of Chinese-Tibetan machine-aided translation system with higher recall of multi-word equivalent pair.

Keywords: Tibetan information processing; multi-word units; collocation

1 引言

长尾真（Nagao, M.）[1]提出：人类的过程是首先将输入句子分解为片段，接着把这些片断译成目标语言，最后把这些片段合并成长句，其中每个片段采取类比的原则进行翻译。这些片段可以是词、短语或其他由多个词组合而成的语言单位，我们将这些语言单位统称为多词单元。多词单元是单词的扩展，单词和多词单元一起构成了翻译的基本单位。在汉藏翻译过程中，从翻译人员的实践来看，仅仅把词作为翻译的基本单位并不合适，将多词单元作为一个整体来翻译更能够保证译文的准确度和流利度，这种整体性的翻译对于提高全文翻译的质量是大有好处的。

本文将要构建汉藏辅助翻译系统的多词单元翻译词典，其中每条记录包含汉语有效多词单元以及对应的藏文译文。基于双语语料库进行翻译词典编纂，国内外很多研究者都做了大量工作[2-3]。在汉藏短语对抽取方面，国内已经有些研究。文献[4]提出藏文词串频率统计算法（简称 TSM）和藏文词串序列相交算法（简称 TIA）两种方法进行汉藏短语对抽取。TSM 算法使用藏文词序列相交短语译文获取模型（Sequence Intersection Based Phrase Translation Extraction Model，简称 SIBPTM 模型），对句对齐双语语料库中包含待翻译汉语语块的句对集合求交集来抽取译文。为了提高准确率，SIBPTM 模型以汉藏词典为辅助资源，并设定阈值解决部分未登录现象。由于使用的汉藏双语词典覆盖率较低，未登录现象较突出，所以，这种方法能够抽取的短语对规模有限。

* 本文承中国科学院西部行动计划高新技术项目(KGCX2-YW-512)的资助。

如果用大规模语料库进行训练以扩大覆盖率，一定程度上可以弥补召回率低的缺陷，但是汉藏机器翻译的研究还处于起步阶段，平行语料库规模十分有限。因此，在当前形势下，相对而言，准确率显得不是特别重要，如何提高召回率是当前更需要考虑的问题。

本文重点研究如何提高基于汉藏对齐语料库的多词单元等价对抽取方法召回率的问题。

2 基于关联度的多词单元等价对获取模型

本文提出 CMWEPM（Collocation Based Multi-Word Equivalence Pair Extraction Model）模型来抽取汉藏多词单元等价对。与文献[4]中 SIBPTM 模型类似，CMWEPM 模型分两步完成等价对抽取，但是它在获取有效汉语语块及确定译文方法上均与前者不同。

为了识别汉语多词单元，本文使用 Ying Zhang 和 Ralf Brown 等人[5]提出的关联度（*Collocation*）度量指标。下面简要介绍这个度量指标。

2.1 关联度

Collocation 可以比较全面地衡量事件关联度，其定义如下：

$$Collocation(w_1, w_2) = \frac{VMI(w_1, w_2)}{H(w_1) + H(w_2)} \tag{1}$$

其中，*VMI* 是平均互信息；w_1, w_2是待衡量的两个单词的出现。*VMI*定义如下：

$$\begin{aligned} VMI(w_1, w_2) = {} & P(w_1, w_2)\log\frac{P(w_1, w_2)}{P(w_1)P(w_2)} + P(\overline{w_1}, \overline{w_2})\log\frac{P(\overline{w_1}, \overline{w_2})}{P(\overline{w_1})P(\overline{w_2})} \\ & - P(w_1, \overline{w_2})\log\frac{P(w_1, \overline{w_2})}{P(w_1)P(\overline{w_2})} - P(\overline{w_1}, w_2)\log\frac{P(\overline{w_1}, w_2)}{P(\overline{w_1})P(w_2)} \end{aligned} \tag{2}$$

其中，$P(w_1, w_2)$是 w_1为第一个单词、w_2为第二个单词，在原文句子中相邻出现的概率；$P(\overline{w_1}, \overline{w_2})$是 w_1 和 w_2 都不出现的句子概率；$P(w_1, \overline{w_2})$ 是 w_1 出现，w_2 不出现的句子概率。

H 指每个词所含的信息量的统计。本文将计算的汉语单词的平均信息量定义如公式（3）。

$$H(w) = -[P(w)\lg P(w) + P(\overline{w})\lg P(\overline{w})] \tag{3}$$

可以看出，在 *VMI* 的计算公式中，前两项表现了对两个词共现有贡献的互信息；后两项表现了对共现有抵消作用的互信息。平均互信息能够综合考虑整个语料库情况，可以全面地衡量两个词之间的关联度。

然而，平均互信息值也只是说明了两个词共现的趋势大小，该值高只能表明 w_1 、w_2 同时出现的趋势大，可能它们其中一个或者两个都是高频词，因此，这两个词出现的频率应该被考虑进去。式中分母即是 w_1 、w_2 的平均信息量，对平均互信息值起到归一化的作用。

假设句子片段包含三个词 w_1, w_2, w_3. 将 w_1 与 w_2 的 *Collocation* 值记为 x, w_2 与 w_3 的 *Collocation* 值记为 y,则 *BindingDegree*(x, y) 计算方法如下：

$$BindingDegree(x, y) = \begin{cases} x / y, & \mathrm{y} \geqslant \mathrm{x} \\ y / x, & \mathrm{y} < \mathrm{x} \end{cases} \tag{4}$$

在这里，*BindingDegree*(x, y) 用于衡量多词单元中词语的结合度并确定多词单元的边界。以下将 *BindingDegree*(x, y) 称为结合度，它计算出的值简称 *BD* 值。

2.2 约束多词单元

CMWEPM 模型是基于词对齐的，利用关联度和结合度确定汉语多词单元边界后，通过词对齐优化结果选择汉语多词单元的译文。利用 GIZA++获得的词对齐矩阵是等价对抽取的起点。

Koehn[6]提出了基于词对齐的完全相容的短语翻译模型。下面先给出短语定义。设: $f=f_1\cdots f_m, e=e_1\cdots e_n$ 分别为源语言和目标语言句子，α 是两个句子上的对齐，则短语互译对 $<e_{i1}\cdots e_{im}, f_{j1}\cdots f_{jn}>$ 是与 α 一致的，当且仅当有下列条件成立:

(1) $\forall j \neg\exists i'(i',j')\in\alpha, i'\notin\{i_1,\cdots,i_m\}, j\in\{j_1,\cdots,j_n\}$；

(2) $\forall i \neg\exists j'(i',j')\in\alpha, i\in\{i_1,\cdots,i_m\}, j'\notin\{j_1,\cdots,j_n\}$；

(3) $\exists k,l(i_k,j_l)\in\alpha, 1\leqslant k\leqslant m, 1\leqslant l\leqslant n$.

Koehn 抽取方法是严格按照词对齐进行的，因此本文称此类多词单元为严格多词单元。它要求完全相容，因此抗噪声能力不强。本文将从汉藏多词单元等价对抽取实际出发，放宽一致性的条件，采用基于词汇相似度约束的抽取策略来减少错误词对齐结果造成的精度损失。能够满足公式(5)中对齐约束条件的汉藏多词单元等价对均被抽取，从而避免破坏等价对的完整性。

$$\forall(i,j)\in\{(i,j)\mid sim(e_i,f_j)\geqslant\theta,(i,j)\in\alpha\},$$
$$((i\in\{i_1,\cdots,i_m\}\wedge j\in\{j_1,\cdots,j_n\})\vee(i\notin\{i_1,\cdots,i_m\}\wedge j\notin\{j_1,\cdots,j_n\})) \tag{5}$$

满足公式(5)的词串为约束多词单元，其中，$\mathrm{sim}(e_i,f_j)$ 是词汇相似度度量函数，θ 是阈值。

2.3 多词单元分类与阈值选取

对于高频多词单元和低频多词单元设定同一个阈值并不合理，本文应用四点法则弱化主观影响且不失多词单元的全面性，从而降低阈值对精确度的影响，提高准确度和效率。为了使计算更有针对性，本文将多词单元分为以下四类：（1）短高频多词单元；（2）短低频多词单元；（3）长高频多词单元；（4）长低频多词单元。表 1 给出类型趋向与关联度和结合度对应情况。

设定四种阈值与多词单元类型对应，保证阈值的选取对多词单元类型具有更好的分辨力。阈值选取以关联度和提取出的多词单元的长度作为参考因素，基本上权衡这两方面就可以。约定横坐标表示 *Collocation* 值，纵坐标表示 *BindingDegree* 值；本文实验所使用的短高频、短低频、长高频和长低频对应的一组参考阈值用坐标形式表示如下：A(0.38，0.6)，B(0.1，0.6)，C(0.38，0.3)，D(0.1，0.3)；其中 *Collocation* 和 *BindingDegree* 的高值和低值的阈值分别设定为 thresh_col1=0.38、thresh_col2＝0.1、thresh_sim1＝0.3、thresh_sim2＝0.6。需要说明的是，这些值都无须非常精确，只要结果大体符合以上分类的标准就可以，在后面的处理中还会有进一步的调整。

表 1 多词单元分类表

类别 \ 度量指标	关联度	结合度
短高频多词单元	高	高
短低频多词单元	低	高
长高频多词单元	高	低
长低频多词单元	低	低

2.4 识别汉藏多词单元等价对实例

本节举例说明 CMWEPM 模型提取多词单元等价对的流程。首先，预处理双语语料；得到的汉藏句对如图 1，分词后的汉语和藏文句子分别用 CS 和 TS 表示，句子中的词用空格隔开。

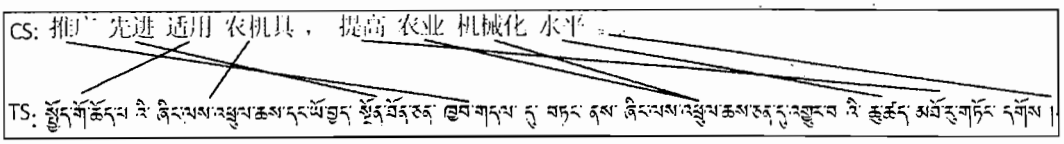

图 1 实例词对齐结果

第二步，计算汉语多词单元。图 2 给出 CS 中相邻词的关联度计算结果。

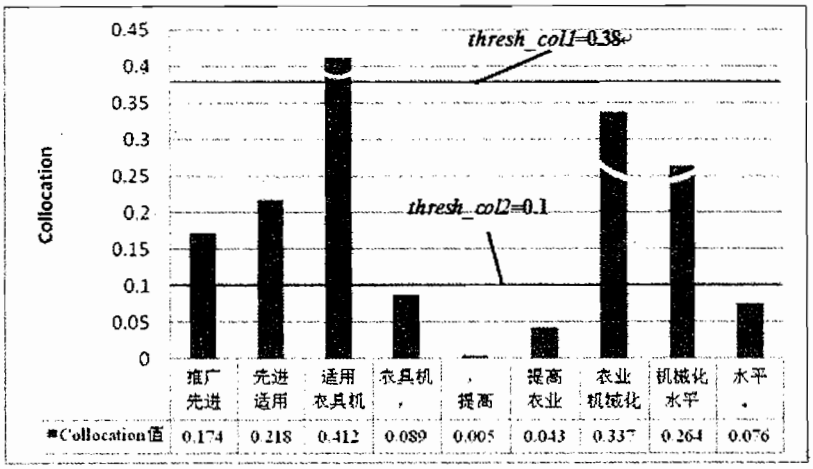

图 2　例句关联度直方图

在图 2 中，“提高”和“农业”的 *Collocation* 值 0.043，小于阈值 thresh_col1=0.38；因此“提高”和“农业”不是多词单元。“农业”与“机械化”的 *Collocation* 值 0.337，大于阈值 thresh_col2=0.1；“机械化”与“水平”的 *Collocation* 值 0.264，这两个关联度的 BD (0.264/0.337)=0.783，大于阈值 thresh_sim1=0.3；“水平”与“。”的 *Collocation* 值 0.076，BD (0.076/0.264)=0.288；小于 thresh_sim1=0.3；因此“农业”、“机械化”和“水平”是一个长低频多词单元。依此类推，“适用”和“农机具”是个短高频多词单元。用‘//’号分割的 CS 的多词单元划分结果如下。

CS 多词单元划分：推广 // 先进 // 适用 农机具 //， // 提高 // 农业 机械化 水平//。//

第三步，应用 Giza++得到词对齐结果。图 1 表示 CS 与 TS 词对齐信息：1-5 2-4 3-1 4-3 6-12 7-9 8-9 9-11 10-14。最后，利用词对齐结果和汉语多词单元划分结果，获取严格的或约束的翻译等价对。可以确定两组汉到藏的多词单元等价对：“适用 农具机，སྤྱད་ཚོད་པ་ ཞི་ཞིང་ལས་འཕྲལ་ཆས་དང་ཡོ་བྱད་”，“农业 机械化 水平，ཞིང་ལས་འཕྲལ་ཆས་ཅན་དུ་འགྱུར་བ་ རྒྱུན་”。

3　实验

文献[4]提出的 SIBPTM 模型和本文提出的 CMWEPM 模型抽取汉藏多词单元等价对的流程都是先抽取汉语有效语块，二者的不同之处在于确定汉语语块边界及获取藏语译文过程。本文将比较两个模型抽取效果，证明本文提出的 CMWEPM 模型的有效性。

在实验中，SIBPTM 和 CMWEPM 两个模型对两组语料进行多词单元等价对抽取之后，采用人工抽样检查的方法判断互译对正确与否，实验的准确率(P)定义为：$P = N / N_t$；召回率(R)定义为：$R = N / N_a$。其中，N 是算法抽取到的正确的多词单元对数目，N_t 是算法从语料库中抽取到的多词单元对数目，N_a 是人工从平行语料中抽取到的多词单元等价对数目。将 P 和 R 两个指标综合为二者的调和平均值 F-Score 来反映一个系统的整体性能：F = 2PR/ (P+R)。

3.1　语料与汉语多词单元规模

实验所采用的双语语料库为表 2 给出的两组语料，其内容主要是汉藏法律法规和公文报告等特定领域语料。语料 1 包括 7 万余已经对齐的汉藏句子。为了提高词对齐准确度，将双语词典加到对齐语料中，最终获取 27 万余对齐句对语料 2。SIBPTM 模型和 CMWEPM 模型抽取汉语多词单元结果统计见表 3。

表2 语料信息

语料	汉语（KB）	藏文（KB）	句对数
语料1	4934	19319	70378
语料2	6959	27584	271416

表3 汉语多词单元信息

语料	汉语文档（KB）	汉语多词单元数量
SIBPTM 模型	51	2206
CMWEPM 模型	338	15052

3.2 多词单元抽取

SIBPTM 模型抽取的多词单元对结果见如表 4。为了提高准确率 TIA 依赖汉藏词典作为辅助资源。由于自然语言翻译的灵活性和双语词典的有限性，词典译项对真实文本的覆盖率很低，SIBPTM 模型进行机械匹配来筛选汉藏多词单元，未登录现象严重影响抽取到的多词单元数量，导致召回率过低。本文尝试用 CMWEPM 模型先获取汉语多词单元，再采取两种策略抽取基于 Giza++词对齐结果的译文。用 Koehn 抽取方法获取的严格多词单元对结果见表 5。已获取汉语有效多词单元和词对齐信息的基础上，本文应用 Phi 平方系数方法计算词汇相似度约束条件，进而抽取约束多词单元等价对；结果在表 6 中给出。表 4-6 中，C1 表示语料 1，C2 表示语料 2。

表4 SIBPTM 模型抽取结果

	P（%）	R（%）	F（%）
C1	82.14	11.93	20.83
C2	83.99	12.20	21.31

表5 严格多词单元抽取结果

	P（%）	R（%）	F（%）
C1	86.44	81.69	84.00
C2	87.81	82.99	85.33

表6 约束多词单元抽取结果

	P（%）	R（%）	F（%）
C1	89.06	84.17	86.54
C2	91.37	86.35	88.79

汉藏语料中数据稀疏问题十分突出。SIBPTM 模型用 n-gram 统计算法抽取汉语多词单元，为了避免太多的干扰信息，过滤掉频次少于 8 的所有多词单元。因此 SIBPTM 模型抽取的汉语多词单元数量很少，这也是下一步实验中造成此模型召回率低的主要原因。将表 5 的结果与表 4 比较可以看出，CMWEPM 模型的召回率比 SIBPTM 模型有明显提高。CMWEPM 模型不再依赖汉藏词典，避免了因词典覆盖率低带来的问题，所以能够提高召回率。并且，CMWEPM 模型使用了成熟的开源词对齐工具进行汉藏词对齐，抽取的多词单元准确率较高。表 6 中数据表明，与严格多词单元结果相比，约束条件的召回率有所提高，这对于处理汉藏语料库有着十分重要的意义。

实验中语料 1 和语料 2 之间的对比表明，扩充训练语料规模能提高 CMWEPM 模型的效能。

4 结语

为了提高汉藏多词单元等价对召回率，本文提出了 CMWEPM 模型。该模型应用关联度和结合度抽取源语言的多词单元，并定义严格条件和约束条件，当词对齐结果符合条件时抽取出此多词单元等价对。实验结果表明，新模型在未经分析语言特征的前提下，取得了令人满意的正确率。与 SIBPTM 模型相比，新模型明显提高了召回率。这对于处理汉藏语料库有着十分重要的意义。

由于藏文形态变化丰富，并且汉语、藏语两种语言差异很大，下一步的工作将考虑加入形态学信息来优化词对齐的准确率，抽取出更为合理的汉藏多词单元等价对。为已经获取的等价对计算翻译概率，应用这些等价对进行翻译解码也是论文下一步工作之一。

参 考 文 献

[1] Nagao, M. A framework of a mechanical translation between Japanese and English by analogy principle[C]//In: Proc. of the international NATO symposium on Artificial and human intelligence, New York, USA, 1984: 173-180.

[2] Jörg Tiedemann. Automatical Lexicon Extraction from Aligned Bilingual Corpora [D]. Magdeburg University, Department of Computer Science, 1997.

[3] 常宝宝. 基于汉英双语语料库的翻译等价单位自动获取研究[J]. 术语标准化与信息技术. 2002. (2): 24-29.

[4] 诺明花, 吴健, 刘汇丹, 丁治明. 汉藏短语对抽取中短语译文获取方法研究[J]. 中文信息学报. 2011. 25(2): 105-110.

[5] Ying Zhang, Ralf Brown, Robert Frederking, Alon Lavie. Pre-processing of Bilingual Corpora for Mandarin-English EBMT[C]//In: proceedings of the MT Summit 8. Santinago de Compostela, Spain, 2001.

[6] Koehn P, Och F J, Marcu D. Statistical phrase based translation[C]// In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown NJ: Association for Computational Linguistics, 2003: 48-54.