文章编号：1003-0077(2015)01-0067-08

高正确率的双语语块对齐算法研究

俞敬松[1,3] 王惠临[2]，吴胜兰[3]

（1. 北京大学 信息管理系，北京 100871；
2. 中国科学技术信息研究所，北京 100038；
3. 北京大学 软件与微电子学院，北京 100871）

摘 要：高质量的自动对齐双语语块，对于机器翻译系统，特别是计算机辅助翻译系统的性能提高有重要作用，而且对于人工翻译以及辞典编纂也都有巨大的应用价值。该文提出基于单词间粘合度与松弛度的语块划分评分方法以及双语语块划分的双向约束算法，使得源语言和目标语言的语块的划分与对齐能相互促进。与传统方法相比，因为无需事先进行双语语块划分，而是在搜索最佳对齐时动态地考察划分效果，故可以减少边界划分错误对对齐结果的影响。该算法获得了远超过传统算法的高正确率。

关键词：语块对齐；机器翻译；平行文本；双语对齐

中图分类号：TP391 **文献标识码**：A

A Bilingual Chunk Alignment Algorithm for Computer Aided Translation

YU Jingsong[1,3], WANG Huilin[2], WU Shenglan[3]

(1. Department of Information Management, Peking University, Beijing 100871;
2. Institute of Scientific and Technical Information of China, Peking University, Beijing 100038;
3. School of Software and Microelectronics, Peking University, Beijing 100871)

Abstract: Automatic Bilingual Chunk Alignment has important application value for Machine Translation, Computer Aided Translation and other fields. In this paper, a Chunk Partition Scoring method is proposed based on the Degree of Adhesion and the Degree of Relaxation to make the chunk partition of source language and target language benefit each other. A novel bilingual chunk alignment algorithm is proposed. Compared with previouswork, this algorithm does not require bilingual chunk partitions, however, the chunk partition score is dynamically calculated during alignment searching. The importance of precision is far beyond recall of this approach.

Key words: chunk alignment; machine translation; parallel corpus; bitext alignment

1 概述

英语语块(chunk)的概念最早由 Abney[1] 提出，代表句子的非递归核心成分，具有句法相关性和不可嵌套性。周强等[2]提出的汉语句子组块分析体系可能是最早的关于汉语语块的研究。对齐的双语语块是在机器翻译研究工作中发展起来的语块扩展形式，程葳等[3]认为双语语块有同样的语义，在翻译上可以互相转换。

双语语块对齐任务可以描述为：给定输入双语句子，自动进行语块划分并按语义对齐。目前多数算法都是基于统计方法的，输出的对齐概率结果面向机器翻译，人工无法解读。

语块对齐工作缺乏标准规范[4]，也没有公开的大规模的标准训练和评测数据。单语语块都很难严谨定义，双语环境中更难。我们认为语块划分要兼顾对齐。从根本来说，语言现象的复杂性是最大的困扰。翻译过程中存在的大量省译、增译、语序调整、意译、兼类、指代等现象加大了双语对齐难度。

我们提出的高质量语块互译对齐，要求系统输出的是人可辨识的有意义结果。译员们获得的是来自计算机的准确的有意义的提示，降低专业译员们的认知负担。在本篇论文中判定是否是语块，除了

收稿日期：2014-04-28 定稿日期：2014-06-16

形式化规则外，主要以人的主观判定为依据：首先语块必须有明确的意义；其次在其他语句中可以重复使用。符合这两条就认为语块划分且对齐正确，没有遵从任何预定义的语法体系。这一点上，本文与其他论文有较大的不同。

本文工作服务于交互式机器翻译等场合[5]：当人类译员在输入完成一个句子的时候，系统依据原文、机器翻译的假设及概率、目标语言模型等进行可能的提示，译员可判断接受从而加速正确译文的产出速度。这里的译文是人的工作成果，与机器翻译没有可比性。高质量双语语块库作为语言资源之一加入模型体系中，提高译员接受猜测的概率。侧重高正确率的算法将依赖更大规模的语料来保证召回。

高质量语块对齐结果对于机器翻译系统来说也是高价值资源。基于短语的机器翻译系统中，过长的句子由于训练时间太耗时而常常被丢弃，利用高质量对齐语块将长句子“拆解”为较短的互译片段可减少训练时间并充分利用语料。本文的语块对齐工作具有语言中立性。

关于语块的研究早期多使用规则方法。吕学强等[6-7]总结了链语法的连接因子和 E-Chunk 的对应关系；刘冬明等[8]在实词对齐的基础上划分语块；屈刚[9]尝试了基于句法深层结构翻译不变性的翻译等价对抽取；Macken[10]的工作则利用了短语构成的语言学知识。近期则是统计方法占主流。姜柄圭等[11]是统计方法为基础，综合运用规则方法抽取语块；刘海霞等[12]将既有的语义资源引入计算过程中；诺明花等[13-14]针对汉藏翻译中藏语语料库规模小，giza＋＋对齐结果不可靠的情况进行了探索；Deng[15]提出了语块对齐的生成模型，本文的工作与之公布的 MTTK 开源系统进行了对比；Liu 等[16]使用 PLSI 技术计算双语语块相似度；Zhao[17]尝试了基于谱聚类的双语词聚类，并基于隐概念流的思想提出了翻译等价对同步生成的图模型；Ma[18]使用了将若干连续的词打包成一个词后反复迭代的方法；Kim[19]将对数线性模型框架应用到了语块对齐任务。

本文在分析和总结前人工作的基础上，创新提出了粘合度与松弛度的概念，分别衡量语块划分时单词间连接紧密程度和松散程度，针对语块划分及对齐提出了多种平行的计算模型及融合算法，并进一步提出了语块对齐时的搜索及扩展算法。

2 语块切分的数学模型的建立

2.1 语块评分方法

给定一个由 N 个词组成的句子 $S = w_1\ w_2 \cdots w_i \cdots w_{N-1}\ w_N$，将相邻词对 $w_i w_{i+1}$ 之间的间隙记作切分点 g_i，其取值 1、0 分别代表划分语块时在该处是否切开。于是，我们可以使用切分点的取值序列 $G = g_1\ g_2 \cdots g_i \cdots g_{N-2}\ g_{N-1}$ 代表语块划分。本文为每个切分点 g_i 设置粘合度与松弛度两个属性（取值均为正整数）：粘合度代表 $w_i w_{i+1}$ 之间的连接紧密程度，记作 a_i，值越大代表连接越紧密；相反地，松弛度代表相邻词对 $w_i w_{i+1}$ 之间的连接的松弛程度，记作 r_i，值越大代表连接越不紧密。假设已知 $w_i w_{i+1}$ 之间粘合度 a_i 和松弛度 r_i，语块划分时在切分点 g_i 处应该切开的概率近似为式(1)。

$$P(g_i = 1) \approx \frac{r_i}{a_i + r_i} \tag{1}$$

若切分点 g_i 处不应该切开，即 $w_i w_{i+1}$ 被划分到同一个语块中的概率近似为式(2)。

$$P(g_i = 0) \approx \frac{a_i}{a_i + r_i} = 1 - P(g_i = 1) \tag{2}$$

请注意一点，黏合度和松弛度是独立计算的。不同的计算模型有不同诉求，可能对两者的计算都有贡献，也可能只对其中之一有帮助。对句子经过语块划分后，形成若干语块。对于一个包含 M 个单词的语块 $C = w_{i+1}\ w_{i+2} \cdots w_{i+m} \cdots w_{i+M-1}\ w_{i+M}$，其左侧和右侧的切分点 g_i 和 g_{i+M} 取值为 1，而内部所有切分点取值为 0，本文使用式(3)来给 C 在语块划分结果中的好坏程度评分：

$$F_{Chunk}(C) = P_{LM}(C)\ P_{Len}(M) P(g_i = 1) \cdot P(g_{i+M} = 1) \prod_{j=i+1}^{i+M-1} P(g_j = 0) \tag{3}$$

其中 $P_{LM}(C)$ 为构成 C 的词语序列在语言模型中的概率，反映语块的频性特征；代表一个语块包含单词数是 M 的先验概率是 $P_{Len}(M)$，反映语块长度的合理性。这个概率分布可以从人工标注语料库统计得到；$P(g_i = 1)$ 和 $P(g_{i+M} = 1)$ 分别代表语块边界上的首尾切分点切开的概率，$\prod_{j=i+1}^{i+M-1} P(g_j = 0)$ 代表语块内部各个切分点是保持连接的。总概率反映语块内部具有一定的句法结构。实际计算时，或可用对数和代替。

2.2　黏合度与松弛度计算方法

本文使用了多种模型分别计算黏合度和松弛度，包括多种自然语言问题的研究成果，例如，依存句法分析，特定成分识别（如时间表达、命名实体）等，其中双语划分镜面约束是利用 GIZA＋＋词语对齐结果矩阵让互译句子对的语块划分互相约束、互相促进，这是其他人的工作中没有讨论过的创新做法。

由于每一种特征模型都可以给出每一个词间切分点 g_i 黏合度和松弛度，由此构成了一个巨大的状态空间，我们的任务就变成了对每种特征权重进行寻优，然后进行融合。在本文工作中，判定经验公式定义的合理与否及取值方法，还有不同影响因素之间的权重关系，均主要靠参数寻优过程来解决。

以命名实体识别特征模型为例，假如句子 S 内部有包含 M 个单词的连续单词串 $w_i\ w_{i+1}\cdots\ w_{i+M-2}\ w_{i+M-1}$ 构成一个命名实体，如人名、地名、机构名等，使用公式(4)计算受影响的各个切分点的黏合度，a_i，λ，θ 为比例因子。

$$a'_k = a_k + W_{NE}\cdot\lambda,\quad k\in\{i,i+1,\cdots,i+M-2\} \tag{4}$$

其中 W_{NE} 为命名实体权重，使用公式(5)计算各个切分点的松弛度 r_i。

$$r'_k = r_k + W_{NE}\cdot\theta,\quad k\in\{i-1,i+M-1\}-\{0,N\} \tag{5}$$

2.3　语块划分评分方法

假设包含 N 个单词的句子 S 经过语块划分（用 G 表示）得到了 $K(1\leqslant K\leqslant N)$ 个语块 $S=C_1\ C_2\cdots\ C_k\cdots\ C_{K-1}\ C_K$，使用公式(6)评估语块划分 G 的好坏程度。

$$\mathrm{F}_{Split}(\mathrm{G},\mathrm{K},\mathrm{S},\mathrm{N}) = \mathrm{P}(\mathrm{K}|\mathrm{S},\mathrm{N})\sum_{k=1}^{K}\mathrm{F}_{Chunk}(C_k) \tag{6}$$

其中，P(K|S,N)为包含 N 个单词的句子 S 最终划分得到的语块总个数是 K 的先验概率，$\mathrm{F}_{Chunk}(C_k)$ 为第 k 个语块 C_k 的评分。句子内容虽可无限变化，句子内单词个数取值分布有规律可循，语块个数在大量数据统计意义下也与单词个数相关，我们认为 $\mathrm{P}(\mathrm{K}|\mathrm{S},\mathrm{N})\approx\mathrm{P}(\mathrm{K}|\mathrm{N})$，概率分布 $\mathrm{P}(\mathrm{K}|\mathrm{N})$ 通过使用人工标注语料库统计结果近似。

2.4　语块互译对相似度计算方法

我们使用式(7)评估中文语块 $\mathrm{X}=c_{i+1}\ c_{i+2}\cdots\ c_{i+m}\cdots\ c_{i+M-1}\ c_{i+M}$ 和英文语块 $\mathrm{Y}=\mathrm{e_{j+1}\ e_{j+2}}\cdots\mathrm{e_{j+u}}\cdots\mathrm{e_{j+U-1}\ e_{j+U}}$ 构成一对语块互译对 $|\mathrm{X},\mathrm{Y}|$ 好坏程度。

$$\mathrm{Score}(|\mathrm{X},\mathrm{Y}|) = \mathrm{P}_{Mode}(|\mathrm{X}|,|\mathrm{Y}|)\cdot \mathrm{P}_{c2e}(\mathrm{Y}|\mathrm{X})\mathrm{P}_{e2c}(\mathrm{X}|\mathrm{Y}) \tag{7}$$

其中，$\mathrm{P}_{Mode}(|\mathrm{X}|,|\mathrm{Y}|)$ 为语块长度（包含的单词数）对齐模式的先验概率，同样从人工标注对齐语料库统计结果中近似。$\mathrm{P}_{c2e}(\mathrm{Y}|\mathrm{X})$ 代表中文语块 X 翻译为英文语块 Y 的概率，$\mathrm{P}_{e2c}(\mathrm{X}|\mathrm{Y})$ 代表英文语块 Y 翻译为中文语块 X 的概率，它们的计算公式为式(8)～(9)。

$$P_{c2e}(Y\mid X)\approx\left(\prod_{m=1}^{M}\max_{u\in\{1\cdots U\}}(\max(P(e_{j+u}\mid c_{i+m}),P(\phi\mid c_{i+m})))\right)^{\frac{1}{M\cdot W}} \tag{8}$$

$$P_{e2c}(X\mid Y)\approx\left(\prod_{u=1}^{U}\max_{m\in\{1\cdots M\}}(\max(P(c_{i+m}\mid e_{j+u}),P(\phi\mid e_{j+u})))\right)^{\frac{1}{U\cdot W}} \tag{9}$$

其中 W 代表 GIZA＋＋词语对齐矩阵中在语块互译对内部的互译词个数，$P(e_{j+u}\mid c_{i+m})$ 和 $P(\phi\mid c_{i+m})$ 分别代表中文单词 c_{i+m} 翻译为英文单词 e_{j+u} 的概率和翻译为空的概率，$P(c_{i+m}\mid e_{j+u})$ 和 $P(\phi\mid e_{j+u})$ 分别代表英文单词 e_{j+u} 翻译为中文单词 c_{i+m} 的概率和翻译为空的概率，这几个概率值都是通过 GIZA＋＋词语对齐训练出来的概率词典得到。上面的式子的意义是（以 $\mathrm{P}_{c2e}(\mathrm{Y}|\mathrm{X})$ 为例）：针对中文语块中的每个中文单词，找出它最可能翻译为英文语块中的哪个单词，如果找不到合适，就将这个单词当作空译处理，并累乘相应的单词翻译概率或空译概率（单词翻译为空的概率），最后对累乘得到的总乘积按照中文单词数与互译词个数乘积的倒数求幂做相应的归一化和奖励。

2.5　语块对齐表示方法

给定源语言句子 S 和目标语言句子 T，假设对 S 和 T 完成了语块划分和对齐后，S 由 K 个语块组成，用 $X_1\ X_2\cdots\ X_k\cdots\ X_{K-1}\ X_K$ 表示，其中 X_k 代表 S 的第 k 个语块，T 由 L 个语块组成，用 $Y_1\ Y_2\cdots\ Y_l\cdots\ Y_{L-1}\ Y_L$ 表示，其中 Y_l 代表 T 的第 l 个语块。语块对齐结果可以用一个目标语块偏移向量 $A=A_1\ A_2\cdots\ A_k\cdots\ A_{K-1}\ A_K$ 来表示，向量的长度与源语言语块个数相同，并满足：

1. 如果第 k 个源语言语块有对应的目标语言

语块，则 A_k 代表第 k 个源语言语块与第 A_k 个目标语言语块对齐。

2. 如果第 k 个源语言语块没有对应的目标语言语块，则约定 $A_k = 0$。

3. 如果第 l 个目标语言语块没有对应的源语言语块，则 l 的值不应该出现在对齐向量所有元素的取值集合中，即 $l \notin \{A_k \mid 1 \leqslant k \leqslant K\}$。

图1为一个使用目标语块偏移向量 A 来表示语块对齐结果的例子。

		攻击 X1	中国 X2	汇率 政策 X3	的 X4	政治 及 经济 X5	理由 X6	越来越 充分 X7	。 X8
the politics and economics	Y1					A / B / C			
of	Y2				D				
an assault	Y3	E							
on chinese	Y4		F						
exchange rate policy	Y5			G					
are increasingly convincing	Y6								
。	Y7								
		A1=3	A2=4	A3=5	A4=2	A5=1	A6=0	A7=6	A8=7

图1 一个用语块偏移向量表示语块对齐的例子

上图中，横坐标从左到右为源语言词序列，纵坐标自上而下为目标语言词序列，粗线条形成的单元格划分语块序列 X1，X2，…，X8 以及 Y1，Y2，…，Y7。最下面的一行单元格代表目标语块偏移向量 A1，A2，…，A8，取值代表对齐，A6＝0 代表该列上的语块没有译文语块。我们研发了专门的 Excel 程序，以图1为模版辅助语料的人工标注及修改以及对齐数据的自动回收和展示，其标注效率和用户亲和程度超过其他方案。

2.6 语块对齐评分方法

在源语言句子 S 和目标语言句子 T 的语块划分分别是 $\mathbf{G}_S$、$\mathbf{G}_T$ 的情况下，将语块对齐用目标语块偏移向量 **A** 表示，本文使用公式(10)计算语块对齐 **A** 的分数。

$$F_{Align}(\mathbf{A},\mathbf{G}_S,K,\mathbf{G}_T,\mathrm{L}) = \sum_{\mathrm{k}=1,\mathrm{A_k}\neq 0}^{\mathrm{K}} \mathrm{Score}(\langle \mathrm{X_k},\mathrm{Y_{A_k}}\rangle) + \sum_{\mathrm{k}=1,\mathrm{A_k}=0}^{\mathrm{K}} \mathrm{Score}(\langle \mathrm{X_k},\phi\rangle) + \sum_{\mathrm{l}=1,\mathrm{l}\notin\{\mathrm{A_k}\mid 1\leqslant \mathrm{k}\leqslant \mathrm{K}\}}^{\mathrm{L}} \mathrm{Score}(\langle \phi,\mathrm{Y_l}\rangle) \quad (10)$$

其中 K 和 L 分别代表源语言句子 S 和目标语言句子 T 语块划分后的语块个数，$\sum_{k=1,A_k\neq 0}^{K} Score(\langle X_k, Y_{A_k}\rangle)$ 为源语言语块和目标语言语块都不为空的语块互译对的相似度分数之和，$\sum_{k=1,A_k=0}^{K} Score(\langle X_k, \phi\rangle)$ 和 $\sum_{l=1,l\notin\{A_k\mid 1\leqslant k\leqslant K\}}^{L} Score(\langle \phi, Y_l\rangle)$ 则分别表示目标语言语块为空的和源语言语块为空的两种情况。

由于本文将语块划分与对齐当整体考虑，故将划分与对齐的综合评分函数定义为源语言划分分数、目标语言划分分数、二者对齐分数的乘积，如式(11)所示。

$$F_{All}(\mathbf{A},\mathbf{G}_S,\mathbf{G}_T,\mathrm{K},\mathrm{L},\mathrm{S},\mathrm{M},\mathrm{T},\mathrm{N}) = \mathrm{F}_{Split}(\mathbf{G}_S,\mathrm{K},\mathrm{S},\mathrm{M})\cdot \mathrm{F}_{Split}(\mathbf{G}_T,\mathrm{L},\mathrm{T},\mathrm{N}))\cdot \mathrm{F}_{Align}(\mathbf{A},\mathbf{G}_S,\mathrm{K},\mathbf{G}_T,\mathrm{L}) \quad (11)$$

3 搜索寻优算法

本文提出语块对齐搜索算法关键数据结构是一个限长 N-Best 列表，列表的每一项用以存储当前划分 $\mathbf{G}_S$、$\mathbf{G}_T$，当前对齐 A，以及划分与对齐总分数 F_{All}。向 N-Best 列表插入新的元素时，新插入的元素自动按照总分数 F_{All} 排序，列表饱和时，排名最靠后的元素则被自动剔除。算法流程如下：

1. 构造初始对齐对(3.1节)，计算其总分数，插入 N-Best 列表；

2. 随机选取出候选对齐，修改候选对齐(3.2节)，计算新对齐与候选对齐之间的总分数增益；

3. 若增益大于0，则将新的对齐也插入 N-Best

列表；

4. 若不满足终止条件，转到第 2 步。

算法设置的默认 N-Best 列表长度为 10 000，默认的终止条件为了 N-Best 列表前三名总分数的平均值连续 1 000 次不变。新对齐与候选对齐之间的总分数增益定义为式(12)。

$$Gain(F_{All}) = F_{All}(\mathbf{A}', \mathbf{G}'_S, \mathbf{G}'_T, K', L' \mid S, M, T, N) - F_{All}(\mathbf{A}, \mathbf{G}_S, \mathbf{G}_T, K, L \mid S, M, T, N) \quad (12)$$

其中 $\mathbf{A}', \mathbf{G}'_S, \mathbf{G}'_T, K', L'$ 分别为修改对齐后形成的新的对齐、新的源语言语块划分、新的目标语言语块划分、新的源语言语块个数，新的目标语言语块个数。

3.1　构造初始对齐

GIZA++词语对齐工具[①]能产生两个方向词语对齐结果，对称化处理可以得到合并的词语对齐结果。但在本文工作的语境下，我们期望的是在正确的前提下得到更长的对齐对。图 2 带有“@”符号的单元格矩阵为 GIZA++词语对齐结果。

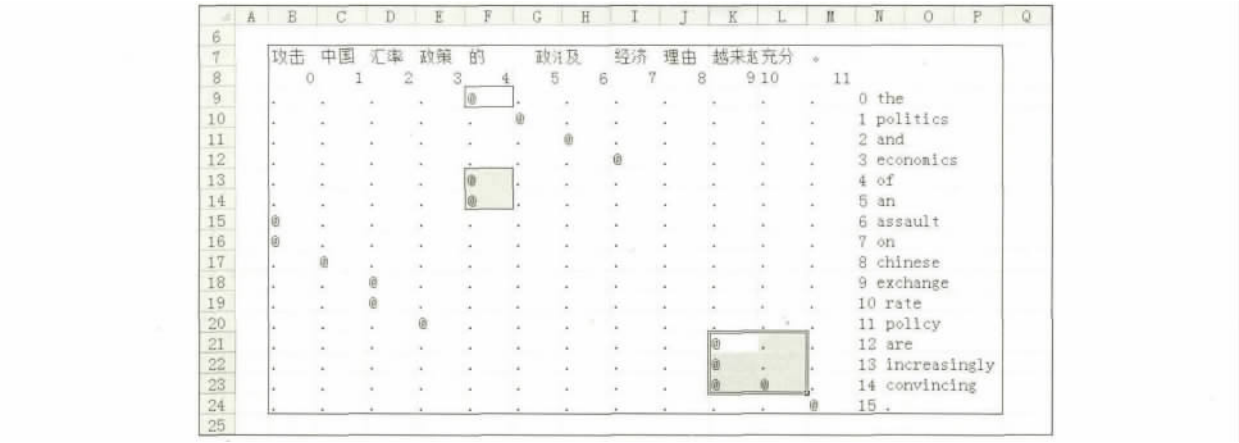

图 2　使用 Excel 作为语块对齐的人工标注和结果检查交互界面

词语对齐矩阵中所有的实词(主要包括名词、动词、形容词)暂可看作互译语块，所有剩余单词看作翻译为空的语块。初始对齐从 GIZA++词语对齐开始，放弃了句子对中意译对齐语块带来的计算量和搜索空间，提高了搜索速度，但代价则是包含意译现象过多的实例中受 GIZA++词语对齐错误影响较大。

本文利用以下规则对不满足语块对齐要求的部分进行冲突消解处理，提高初始语块对齐质量：

1. 词语对齐矩阵中的词在水平或者垂直方向上连续多个对齐时，例如，“assault on”、“exchange rate”，则将其绑定为语块，否则，每个单独的词语独立看待，没有对应的则对空。

2. 对齐冲突时，分别尝试当前源语块与每个目标语块对应，保留相似度最高者，删除的语块看做空译语块。本例中“的”与“the”的词语对齐关系被删除，最终“the”构成空译语块，“的”与 “of an”构成互译对。

3. 对齐矩阵中词对齐单元格形成“+”形、“T”形、“L”形或更复杂的连通区域时绑定对齐，例如，“越来越 充分”与“are increasingly convincing”的词语对齐信息构成“L”形，可构成对齐。

4. 反复重复上述过程，直到不存在冲突。

3.2　修改与扩展候选对齐

在以上计算结果的基础上，首先利用相邻词扩展互译语块，其次合并相邻互译语块。相邻词扩展时，从互译语块的边界出发，向相邻的八个方向扩展，如图 3 所示。

	中国	第一	艘	航母	首	次	出海	试航
China	A	.	.	.	.	.	.	.
has	.	.	.	.	.	.	.	.
launched	.	.	.	.	.	.	.	.
its	.	.	.	.	.	.	.	.
first	.	C	.	.	.	.	.	.
aircraft	.	.	.	.	.	.	.	.
carrier	.	.	.	.	.	.	.	.

图 3　选取互译语块示例

其中“第一”和“first”是一对可以通过查询词典得到的互译实词，初始候选已对齐，其扩展方式有 8 种如图 4。如某方向上有属于其他互译语块的词，则放弃扩展。合并相邻互译语块的方法为随机选择

① http://www.statmt.org/moses/giza/GIZA++.html

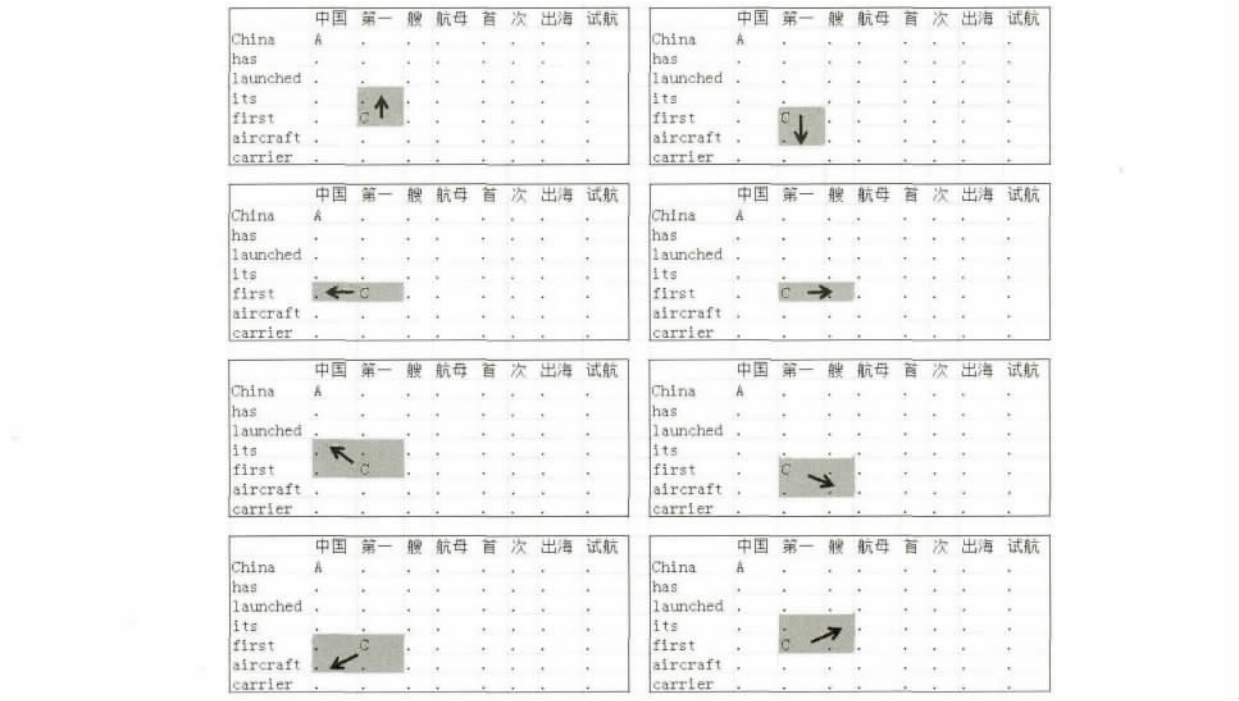

图4 互译语块向周围八个方向扩展示例

两个相邻的互译语块，将二者以及夹在二者之间空译语块合并，如图5所示。

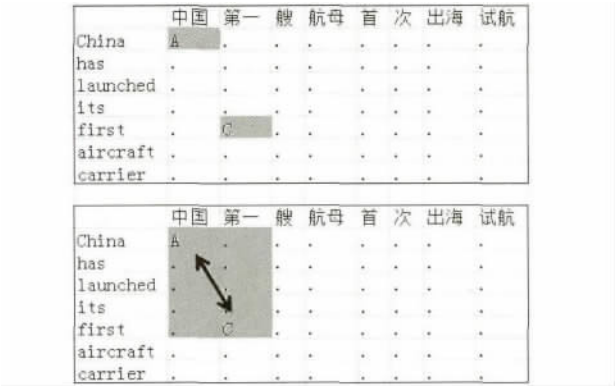

图5 随机合并两个相邻互译语块示例

3.3 减小搜索空间的剪枝策略

为了减小搜索开销，本文进行状态空间剪枝：

1. 设置合适的N-Best列表长度，利用N-Best列表自动删除最差元素的机制过滤；

2. 针对每个候选对齐计算hash，避免重复处理同一个候选对齐；

3. 根据源语言和目标语言句子长度估算出各自的可能语块个数，约束状态扩展的进程；

4. 包含标点符号的语块一般不再与任何语块合并；

5. 优先选择所有候选中增益最大的互译语块以及相应的方向进行扩展；

6. 如果每个方向的增益都小于0，则做标记及下次不再选择。

4 基于人工标注语料库的权重参数优化方法

多种特征模型分别给出了对切分点 g_i 的黏合度和松弛度的独立贡献，融合这些模型需要确定对应的权重参数。我们在人工标注语料库上均匀抽取102个句子对，作为训练集，试图找出一组局部最优的权重参数来优化语块划分的评分值：

（1）根据经验构造一个初始参数集合，用参数序列 $\mathbf{W}=W_1 W_2 \cdots W_N$ 表示，并用另一个N-Best自动排序列表存放最多10 000个最佳参数集合。

（2）利用本文提出的方法计算小规模人工标注的训练集的划分分数的平均值，将初始参数序列W以及对应的平均分数插入N-Best列表。

（3）从N-Best列表随机选择一组参数，并记录这一组参数在列表中的排名Rank，随机选择其中的一个参数进行一定幅度的变化，变化的幅度与Rank呈正相关。即如果这组参数距离N-Best列表的第一名越近，则变化的幅度越小（在局部最优解附近寻优），否则，变化的幅度越大（放弃当前解，跳跃到一个新的解）。

(4) 在变化后的参数集合 W′的基础上计算小规模人工标注的训练集的平均划分分数和对齐总分数，并插入到 N-Best 列表。

(5) 反复重复步骤(3)、(4)，直到 N-Best 列表的前 3 名的平均分数连续若干次不变。

5　实验结果与分析

本文的基础实验主要是在小规模的语块划分和语块对齐的汉英双语人工标注语料库上进行的，语料库共有 509 个句子对，本文将语料库平均地分为包含 102 句对的开发集和包含 407 句对的测试集。本文首先在整个语料库上对三个先验概率分布 $P_{Len}(M)$、$P(K|N)$、$P_{Mode}(|X|,|Y|)$ 进行了统计，然后利用开发集进行了参数优化。参数寻优后，测试集上进行的语块对齐实验结果与人工标注对比，准确率 0.664，召回率 0.714，F 值 0.684，但这个实验的规模有限，仅供调试程序使用。

高质量语块互译对抽取实验以及与其他工作的对比实验则是在包含 150 万句对的训练语料及 1 000 句对的开放测试集上完成的。正确率评测时，从抽取结果中去重后随机采样 5%由封闭语料的标注员依照同样标准审核，其判定标准是：短语必须人工可辨识(有意义)；在翻译其他句子上下文中可复用。表 1 显示互译语块相似度阈值对结果的影响。

表 1　1 000 句对开放测试集正确率抽样评测结果

相似度阈值	非空语块互译对总数	去重后总数	采样数	正确数	正确率/%
−10	4 715	2 583	129	122	94.46
−13	5 811	3 627	181	159	87.68
−50	7 004	4 812	241	191	79.38

从表 1 可以看出，随着阈值的提高，抽取结果的正确率逐渐提高，抽取规模逐渐降低。当阈值设置为−10 时，正确率达到了我们认为在计算机辅助翻译应用领域内可以接受的水平。

本文的工作成果与词到短语对齐工具 MTTK①[15]，以及著名开源机器翻译系统 MOSES② 生成的短语表进行了对比。

MTTK 训练后得到的中文到英文、英文到中文的词到短语对齐结果。GIZA++训练得到双向词到词对齐结果，利用 MOSES 提供的脚本工具进行对称化及短语抽取，对称化算法使用默认设置 grow-diag-final-and，最大短语长度设置为 3。MTTK、MOSES 以及本文输出结果列在表 2 中。

表 2　使用 MTTK、MOSES 以及本文算法在 1 000 句对开放集上的互译对抽取结果统计

对齐模式	去重前			去重后		
	MTTK	MOSES	本文算法	MTTK	MOSES	本文算法
N∶0	0	0	1 987	0	0	1 304
0∶N	0	0	1 265	0	0	780
1∶1	5 345	5 395	4 249	3 180	2 558	2 136
1∶2	2 697	1 898	312	2 499	1 735	268
2∶1	2 918	2 354	426	2 736	2 163	404
1∶3	1 389	891	29	1 375	884	29
3∶1	1 669	1 267	70	1 658	1 261	70
2∶2	**3 984**	**2 951**	**1 075**	**3 889**	**2 864**	**1 070**
2∶3	2 680	1 667	212	2 661	1 660	208
3∶2	3 104	2 154	263	3 089	2 149	259
3∶3	3 207	2 147	147	3 203	2 147	147
其他	0	0	723	0	0	719
总计	26 993	20 724	10 758	24 290	17 421	7 394

结果中 MTTK 互译对最多，本文最少。本文独有 0∶N 和 N∶0 模式的空对齐对；1∶1 模式基本上是词语对齐，本文比 GIZA++少是因为语块合并等因素；723 个语块长度大于 4 的对齐对是语块扩展策略做出的独有贡献。

将本文算法与其他方法对比，主要比较了 2∶2 模式语块对，去重后以 10%的采样率进行随机采样后再加人工判定，结果如表 3 所示。

表 3　本文与其他研究工作 2∶2 模式对齐的短语对比

方法	总互译对数	去重后	采样数	正确数/%	正确率	反向估计
MTTK	3 984	3 889	388	93	23.97	932
MOSES	2 951	2 864	286	85	29.72	851
本文算法	1 075	1 070	107	97	90.65	970

反向估计代表的是根据正确率和过滤后的互译对数，估算出过滤后的互译对中正确互译对的数目，

① http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/

② http://www.statmt.org/moses/

3种方法大致相同，由此也可证明本文工作的正确性。

虽然MTTK和MOSES总抽取的互译对数目远超过本文工作，但正确率较低。实验结果表明本文算法抽取结果少而精，以牺牲召回率为代价提高了准确率。

我们相信，考虑到译员的认知负担的问题，对于实际翻译任务来说，本文的工作对译员的帮助更大，因为过多的杂乱结果会干扰他们的思考。

6 总结

本文的工作存在一些缺点和不足：首先是没有考虑跨越多个词的搭配，也不允许嵌套，认为语块只是连续的单词片段，这是有缺陷的。语块相似度计算函数目前还比较简单，尝试利用词性、GIZA++产生的词类信息等或可得到更好结果。

本文从GIZA++词语对齐作为工作起点，但是语块划分与词语对齐似乎依然可以做到相互促进，引文语块对齐可以帮助缩小词语对齐的边界，而高质量的词语对齐也是产生高质量语块对齐的先决条件。这一点还需要进一步的实证研究。

附注：本文的语块对齐部分成果，10万条语块对齐库已经在数据堂网站免费公开，有需要者可以自行下载使用。(http://www.datatang.com/data/46112)公布的数据是对齐程序的直接输出结果，没有再进行其他加工处理。

参考文献

[1] Abney Steven. "Statistical methods and linguistics." [J]. The balancing act: Combiningsymbolic and statistical approaches to language. 1996: 1-26.

[2] 周强，孙茂松，黄昌宁. 汉语句子的组块分析体系[J]. 计算机学报，1999，22(11)：1158-1165.

[3] 程葳，赵军，徐波，等. 一种面向汉英口语翻译的双语语块处理方法[J]. 中文信息学报，2003，17(2)：21-27.

[4] 李业刚，黄河燕. 汉语组块分析研究综述[J]. 中文信息学报，2013，27(3)：1-8.

[5] Ortiz-Martínez D, Leiva L A, Alabau V, et al. Interactive machine translation using a web-based architecture[C]//Proceedings of the 15th international conference on intelligent user interfaces. ACM, 2010: 423-424.

[6] 吕学强，李清隐，任飞亮，等. 基于统计的汉英法律文献亚句子级对齐[J]. 东北大学学报，2003，24(1)：23-26.

[7] 吕学强，陈文亮，姚天顺. 基于连接文法的双语E-Chunk获取方法[J]. 东北大学学报，2002，23(9)：829-832.

[8] 刘冬明，杨尔弘. 一种新的双语语块对应算法[J]. 电脑开发与应用，2004，17(3)：2-3.

[9] 屈刚. 英汉双语短语对齐[D]. 上海交通大学，2007.

[10] Macken Lieve. Sub-sentential alignment of translational correspondences[D]. Ghent University, 2010.

[11] 姜柄圭，张秦龙，谌贻荣，等. 面向机器辅助翻译的汉语语块自动抽取研究[J]. 中文信息学报，2007，21(1)：9-16.

[12] 刘海霞，黄德根. 语义信息与CRF结合的汉语功能块自动识别[J]. 中文信息学报，2011，25(5)：53-59.

[13] 诺明花，张立强，刘汇丹，等. 汉藏短语抽取[J]. 中文信息学报，2011，25(1)：105-110.

[14] 诺明花，吴健，刘汇丹，等. 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(3)：112-117.

[15] Deng Yonggang. Bitext alignment for statistical machine translation[D]. Johns Hopkins University, 2006.

[16] Liu Feifan, et al. Bilingual chunk alignment based on interactional matching and probabilistic latent semantic indexing[J]. Natural Language Processing IJCNLP 2004. Springer Berlin Heidelberg, 2005: 416-425.

[17] Zhao Bing. Statistical alignment models for translational equivalence[D]. Carnegie Mellon University, 2007.

[18] Ma Yanjun, Nicolas Stroppa, Andy Way. Bootstrapping word alignment via word packing[J]. Annual Meeting-Association for Computational Linguistics. 2007，45(1)：304-311.

[19] Kim Jae Dong. Chunk alignment for Corpus-Based Machine Translation[D]. Carnegie Mellon University, 2012.

(下转第110页)

2013,45(11): 45-49.
[11] Liu Yi, Jin Rong, Yang Liu. Semi-supervised multi-label learning by constrained non-negative matrix factorization[C]//Proceedings of the 21 st National Conference on Artificial Intelligence. Menlo Park: AAAI,2006: 421-426.
[12] Chen Gang, Song Yangqiu, Wang Fei, et al. Semi-supervised multi-label learning by Solving a Sylvester equation [C]//Proceedings of SIAM International Conference on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2008: 410-419.
[13] 姜远,余俏俏,黎铭,等. 一种直推式多标记文档分类方法[J]. 计算机研究与发展,2008,45(11): 1817-1823.
[14] Sun Yuyin, Zhang Yin, Zhou Zhihua. Multi-label learning with weak label[C]//Proceedings of the 24 th AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2010: 593-598.
[15] 孔祥南, 黎铭, 姜远,等. 一种针对弱标记的直推式多标记分类方法[J]. 计算机研究与发展. 2010,47(8):1392-1399.
[16] Xiangnan Kong, Michael K. Ng, Zhou Zhihua. Transductive Multi-label Learning via Label Set Propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2013,25(3): 704-719.
[17] 李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习方法[J]. 计算机研究与发展. 2012,49(6): 1272-1278.
[18] 周志华,王珏. 半监督学习中的协同训练算法[M]. 机器学习及其应用. 北京:清华大学出版社, 2007: 259-275.
[19] 刘杨磊, 梁吉业, 高嘉伟,等. 基于 Tri-training 的半监督多标记学习算法[J]. 智能系统学报. 2013, 8(5):439-445.
[20] Zhou Zhihua,Li Ming. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.
[21] http://mulan.sourceforge.net/datasets.html[OL].
[22] Zhou Zhihua, Zhang Minling, Huang Shengjun, et al. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176:2291-2320.

高嘉伟(1980—),讲师,博士研究生,主要研究领域为机器学习。
E-mail: gjw@sxu.edu.cn

梁吉业(1962—),博士,教授,博士生导师,主要研究领域为机器学习、计算智能、数据挖掘等。
E-mail: ljy@sxu.edu.cn

刘杨磊(1990—),硕士研究生,主要研究领域为机器学习。
E-mail: lyl_super@126.com

(上接第 74 页)

俞敬松(1971—),博士研究生,硕士,副教授,主要研究领域为自然语言处理、计算机辅助翻译、教育技术等。
E-mail: yjs@ss.pku.edu.cn

王惠临(1948—),博士生导师,博士,主要研究领域为信息管理、机器翻译、自然语言处理等。
E-mail: wanghl@istic.ac.cn

吴胜兰(1987—),硕士,主要研究领域为自然语言处理、购物商品推荐等。
E-mail: wslgb2010@qq.com