基于序列相交的短语译文获取

王辰[1] 宋国龙[1] 吴宏林[1] 张俐[1] 刘绍明[2]
[1]东北大学自然语言处理实验室 沈阳 110004
[2]富士施乐公司 日本 神奈川
Email: wangchen@ics.neu.edu.cn

摘要：短语译文获取技术是基于实例机器翻译系统EBMT中的核心技术之一，短语译文获取技术的性能直接影响到EBMT的性能。当前主要的短语译文获取方法过于依赖词对齐结果，只能从词对齐库中得到短语译文结果；有些方法利用句法分析结果，存在代价高、翻译精度低等问题。该文提出了一种基于序列相交的短语译文获取方法，只需要句对齐双语语料库，不需要词对齐、句法分析及词典等资源。实验表明，该方法具有较高的准确率。

关键词：EBMT 短语译文获取 序列相交

Phrase Translation Extraction Based on Sequence Intersection in Bilingual Corpus

WANG Chen[1], SONG Guolong[1], WU Honglin[1], ZHANG Li[1], LIU Shaoming[2]
[1] Natural Language Processing Laboratory, Northeastern University, Shenyang, 110004
[2] FujiXerox Co., Ltd. Kanagawa, Japan
Email: wangchen@ics.neu.edu.cn

Abstract: Phrase translation extraction is one of the key techniques in the Example-Based Machine Translation (EBMT) as its performance affect the performance of EBMT system directly. Currently the main methods of phrase translation extraction depend heavily on word alignment, thus translation can only be accessed from the word alignment corpus. Although in some methods syntax parsing is taken into account, the problems of high cost and low accuracy can't be avoided. This paper proposes a phrase translation extraction method based on sequence intersection only using sentence level aligned bilingual corpus rather than the resources like word alignment, parsing and dictionary. The experiments show this approach achieves high accuracy.

Keywords: EBMT; phrase translation extraction; sequence intersection

1 引言

基于实例的机器翻译 EBMT(Example-Based Machine Translation)的思想最早由日本学者Nagao在1981年提出。EBMT的基本思想是:预先构造由双语对照的翻译实例组成的双语平行语料库，然后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，最后根据该实例的译文构造当前所翻译单元的译文。短语译文获取是EBMT中不可缺少的核心环节之一。

目前已经提出了多种短语对抽取的方法，Marcu[1]给出了一种直接计算短语对列表和相应概率值的方法；Wu[2]使用双语框架(Bracketing)方法来抽取短语。Zhang[3]为双语句对建立一个互信息矩阵，矩阵中的每一个单元格是词对的点式互信息，从这个矩阵中抽取互信息相似的矩形区域即得到短语对，此方法并不要求词对齐，而是充分利用词对的互信息。后来Zhang[4]将短语抽取看作一个句子分割问题，在固定原短语时，寻找目标短语的最优左边界和右边界。Kaji[5]对原句子和目标句子分别进行句法分析，然后按照词对齐结果来提取原子树和目标子树就得到短语对，

该方法依赖于句法分析的结果。Och[6]提出了对齐模板方法，将单词映射到词类中。该方法由于算法简单，容易实现，故而应用较广。何彦青[7]给出了一种基于"松弛尺度的短语抽取方法，对Och的方法进行了修改。刘冬明[8]提出了一种在汉英双语语料库句子对齐的基础上，自动进行汉英名词短语划分和对应的方法。屈刚[9]提出了一种基于有效句型的英汉双语短语对齐方法。吴宏林[10]提出了一种多层次的短语译文获取方法。

上述方法中有些计算复杂度太高，代价很高、难于实现；有些依赖于句法分析或词对齐技术，对资源的要求很高。本文提出了基于序列相交的短语译文获取方法，该方法可在没有词对齐、句法分析和词典的帮助下，在句子级对齐双语语料库中对包含待翻译短语的句对求交集，得到候选译文，然后经过后处理得到短语的翻译译文。基于序列相交的短语译文获取方法摆脱了对词对齐、句法分析及词典的依赖，并且准确率较高，可以作为多策略短语译文获取中的一个模块。

本文剩余部分安排如下:第 2 节介绍本文提出的基于序列相交的短语译文获取方法；第 3 节给出实验结果及分析；第 4 节给出结论与下一步工作。

2 基于序列相交的短语译文获取方法

基于序列相交的短语译文获取方法由基本模型、高频干扰词限制模块、支持度限制模块组成。基本模型从句子级对齐双语语料库中提取高质量的短语翻译对候选并对其进行排序；高频词限制模块解决译文的输出结果中的高频词干扰问题；支持度限制模块控制输出结果的个数。

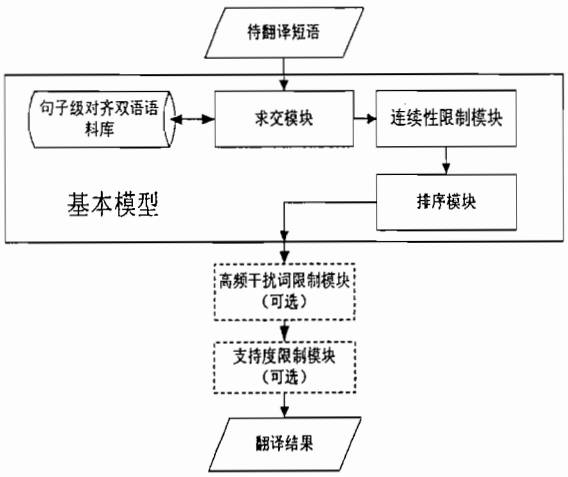

图 1 基于序列相交的短语译文获取方法总体结构图

2.1 基本模型

2.1.1 句子和短语的序列表示

基于序列相交的短语译文获取方法使用的语料库为句子级对齐的双语语料库*BC*，其中包含若干个汉日对齐的句对。句对S表示为$S = CS \leftrightarrow JS$，其中*CS*和*JS*是互为译文的汉语句子和日语句子。在本方法中，句子以字序列的形式表示：

$$CS =< c_1, c_2, \cdots c_m > \tag{1}$$

$$JS =< j_1, j_2 \cdots j_n > \tag{2}$$

这样句对 S 就可以表示成字序列的形式：

$$S = CS \leftrightarrow JS =< c_1, c_2, \cdots c_m > \leftrightarrow < j_1, j_2 \cdots j_n > \tag{3}$$

设 P 为待翻译的中文短语，以字序列的形式表示：

$$P =< p_1, p_2 \cdots p_n > \tag{4}$$

例如：

CS = <中,国,队,面,对,欧,洲,劲,旅,西,班,牙,队,。>

JS = <中,国,チ,一,ム,は,ヨ,一,ロ,ッ,パ,の,強,い,チ,一,ム,ス,ペ,イ,ン,チ,一,ム,に,面,し,て,い,る,。>

$S = CS \leftrightarrow JS$ = <中,国,队,面,对,欧,洲,劲,旅,西,班,牙,队。>↔<中,国,チ,一,ム,は,ヨ,一,ロ,ッ,パ,の,強,い,チ,一,ム,ス,ペ,イ,ン,チ,一,ム,に,面,し,て,い,る,。>

P =< $p_1, p_2 \cdots p_n$ >=<中国队>

2.1.2 句子的序列相交定义

设双语句对 $S_k, S_h \in BC$，

$$S_h = CS_h \leftrightarrow JS_h = < c_h, c_{h+1}, \cdots c_{h+m_h} > \leftrightarrow < j_h, j_{h+1}, \cdots j_{h+n_h} > \tag{5}$$

$$S_k = CS_k \leftrightarrow JS_k = < c_k, c_{k+1}, \cdots c_{k+m_k} > \leftrightarrow < j_k, j_{k+1}, \cdots j_{h+n_k} > \tag{6}$$

Sk 与 Sh 相交定义为：

$$S_h \cap S_k = CS_h \cap CS_k \leftrightarrow JS_h \cap JS_k \tag{7}$$

其中，$CS_h \cap CS_k$ 定义为：

$$\begin{aligned} CS_h \cap CS_k &= \underset{<c_{h+h_1}, c_{h+h_2}, \cdots, c_{h+h_q}>}{\arg\max} \left| < c_{h+h_1}, c_{h+h_2}, \cdots, c_{h+h_q} > \right| \\ &= \underset{<c_{k+k_1}, c_{k+k_2}, \cdots, c_{k+k_r}>}{\arg\max} \left| < c_{k+k_1}, c_{k+k_2}, \cdots, c_{k+k_r} > \right| \end{aligned} \tag{8}$$

$$0 \le h_1 < h_2 < \cdots < h_q \le m_h \tag{9}$$

$$0 \le k_1 < k_2 < \cdots < k_r \le m_k \tag{10}$$

公式(8)表示 $CS_h \cap CS_k$ 的结果为一个新的字序列，该序列中的各个字对应 CS_h, CS_k 中的各单词，两个序列的下标 $h_1, h_2, \cdots h_q$ 和 $k_1, k_2 \cdots k_r$ 应分别落在 CS_h, CS_k 的下标范围内，并为单调递增，即应满足公式(9)和公式(10)。

类似公式(8)，可定义 $JS_h \cap JS_k$ 。

$CS_h \cap CS_k$ 与 $JS_h \cap JS_k$ 举例：

S_k	CS_k	中国队小组出线命运已经不掌握在自己手中。
	JS_k	中国チームは自分の手でグループから勝ち進む運命を握っていなかった。
S_h	CS_h	体力对中国队其实根本不是一个问题。
	JS_h	体力は中国チームにとって、まったく問題にならない。
$CS_h \cap CS_k$	中国队	
$JS_h \cap JS_k$	中国チーム	

2.1.3 基于序列相交的短语译文获取基本模型

若 S_k 与 S_h 的交集为：

$$S_h \cap S_k = P \leftrightarrow T_g = P \leftrightarrow < j_{g_1}, j_{g_2}, \cdots j_{g_m} > \tag{11}$$

P 为待翻译的中文短语，T_g 为 S_k 与 S_h 的日文部分交集。称 S_h、S_k 支持 $P \leftrightarrow T_g$，称 T_g 为 P 的候选译文。若语料库中有 x 个句对支持 $P \leftrightarrow T_g$，则称 T_g 作为 P 的候选译文时的支持度为 x，记为：

$$SV(P \leftrightarrow T_g) = x \tag{12}$$

选择支持度最大的候选译文作为 P 的翻译结果：

$$Translation(P) = \underset{T_g}{\arg\max}\, SV(P \leftrightarrow T_g) \tag{13}$$

若 S_h 与 S_k 的交集为：

$$S_h \cap S_k = P^* \leftrightarrow T_g = P^* \leftrightarrow < j_{g_1}, j_{g_2}, \cdots j_{g_w} > \wedge P \subset P^* \quad (14)$$

P^* 为 S_h 与 S_k 的中文部分交集，P 为待翻译的中文短语，若 P 为 P^* 的连续子串，将 $P^* \leftrightarrow T_g$ 作为新的句对加入 BC，重新与其他句对计算。

相似地，可以定义日文短语的翻译模型。

2.1.3 连续性限制模块

在一般情况下，短语的翻译结果在译文句子中呈现一定的连续性倾向。而基本模型中没有反映这种倾向。

对于 S_h、S_k 支持 $P \leftrightarrow < j_{g_1}, j_{g_2}, \cdots j_{g_w} >$，如果 $g_1, g_2, \ldots, g_{ng}$ 连续，则称 S_h、S_k 强支持 $P \leftrightarrow < j_{g_1}, j_{g_2}, \cdots j_{g_w} >$，称 $< j_{g_1}, j_{g_2}, \cdots j_{g_w} >$ 为 P 的强候选译文；否则，称 S_h、S_k 弱支持 $P \leftrightarrow < j_{g_1}, j_{g_2}, \cdots j_{g_w} >$，称 $< j_{g_1}, j_{g_2}, \cdots j_{g_w} >$ 为 P 的弱候选译文。

在连续性限制模型中，选择支持度最大的强候选译文作为 P 的翻译结果；如果没有强候选译文，则选择支持度最大的弱候选译文作为 P 的翻译结果。

2.2 高频干扰词限制模块

在基本模型中，没有使用词典、词对齐等资源，译文结果是由求交集的结果按照频率排序得到的，因而一些高频词（如：的、地、得、は、が、を等）干扰了译文获取结果。去除排序的候选译文中的这些高频干扰词，即为高频干扰词限制模块。根据在短语译文获取过程中对高频干扰词的限制作用不同把 W^{STOP} 分为 W^{STOP_1} 和 W^{STOP_2}.

2.2.1 高频干扰词限制模块中的 W^{STOP_1}

若待翻译短语 P 的候选译文 T_g 中有 W^{STOP_1} 中的干扰词，即 W^{STOP_1} 中的干扰词作为 P 的候选译文中的一项，并且这个干扰词并不是待译短语对应的译文，那么将这一项去掉，重新根据支持度排序选择最终结果。

2.2.2 高频干扰词限制模块中的 W^{STOP_2}

若待翻译短语 P 的候选译文 T_g 中的某一个候选译文的首尾部分包含 W^{STOP_2} 中的干扰词，那么判断在待翻译短语 P 中是否有这个干扰词对应的待翻译项，若有，就将其保留，否则，将干扰词去除。

2.3 支持度限制模块

由于基本模型没有使用词典、词对齐等资源，无法判断求出的交集结果是否符合译文要求。故当求交结果的支持度很低时，往往得到的不是正确译文；当候选译文之间的支持度比较相近时，只输出一个译文，很可能漏掉正确地译文结果。因此为了提高译文结果的质量，需要一个判定模块，在基本模型中增加对候选译文的支持度的限制，即：支持度限制模块。

设 T_1 和 T_2 分别为 P 的支持度最大的和次大的候选译文。

$$SV(P \leftrightarrow T_1) = x \quad (15)$$

$$SV(P \leftrightarrow T_2) = y \quad (16)$$

IF $x<\theta_1$ 则根据本模块不输出结果；

ELSE IF $x-y>\theta_2$（也可限制为：$x/y>\theta_3$）按基本模型输出结果 T_1；

ELSE 同时输出 T_1 和 T_2 两个结果，在机器翻译中对其进行选择。（$\theta_1, \theta_2, \theta_3$ 为通过实验确定的阈值。）

3 实验

为了验证本文提出的基于序列相交的短语译文获取基本模型，以及高频干扰词限制模块、支持度限制模块对基本模型改进的有效性，本实验中模型使用的句子级对齐的双语语料库 BC 是

10000 句体育领域的中日双语句对齐语料。我们随机抽取了 BC 中的 40000 个中文短语进行测试。

3.1 高频干扰词限制模块的验证

为了验证本文提出来的高频干扰词限制模块的有效性，我们设计了三个系统、两个对比实验：Base_Model 为基本模型；Model_with_W^{STOP_1} 为在 Base_Model 上使用了高频词限制模块 W^{STOP_1} 的模型；Model_with_$W^{STOP_1_WSTOP_2}$ 为在 Base_Model 上同时使用 W^{STOP_1} 、W^{WSTOP_2} 的模型。我们分别对它们的性能进行了测试，并进行了比较。

3.1.1 Base_Model 性能

由于人工评价代价高昂，无法对全部 40000 个短语的译文结果准确率进行评价，故只随机选择了 400 个译文结果，对其正确性进行人工评价，下文出现译文结果准确率评价的地方，都采用了这种随机抽取的方式。译文结果正确的标准为能准确、完整的表达待译短语意思。

表 1 Base_Model 的性能

	Base_Model
译文结果数量	400
正确率	0.845

通过实验结果分析，可以发现通过序列相交方法的得到的译文有着较高的准确率。由于对于本方法的短语译文获取，其译文结果的召回率受到短语覆盖率影响的，很难单纯评价当前待测试短语的召回率，并且其意义并不大，故本实验没有进行召回率的计算。

3.1.2 Model_with_W^{STOP_1} 与 Base_Model 的比较

为了检验 Model_with_W^{STOP_1} 对 Base_Mode 的改进效果，分别对其在 40000 个短语的测试集上进行了测试，得到如下数据。

表 2 Model_with_W^{STOP_1} 与 Base_Model 的比较结果

	Base_Model	Model_with_W^{STOP_1}
有译文结果的短语	12524	12624
改正的错误	--	228
增加的译文结果	--	137
减少的译文结果	--	37

从表中可以看出，在基本模型使用了高频干扰词限制模块的 W^{STOP_1} 后，我们的系统有效的减少了高频干扰词的影响，表现在如下几个方面：

- 改正的错误：最主要的情况为，在待翻译短语 P 的候选译文中，支持度最高的译文是 W^{STOP_1} 中的高频干扰词，并且符合其他限制要求，被当成译文结果输出了，在使用 W^{STOP_1} 后将这一项除去，将次支持度的正确的候选译文项作为输出结果，提高了准确率。例如：

中文待翻译短语 P =“奥运军团”，$P \leftrightarrow T_g$，T_g =<“オリンピック”，“オリンピック代表团”，…>，“オリンピック”的支持度为 35，“オリンピック代表团”的支持度为 10，输出结果应该为“オリンピック”，但是“オリンピック”同样是一个在体育领域的语料库中出现频率很高的短语，属于 W^{STOP_1} 中的高频干扰词，应去掉这项；因此“オリンピック代表团”作为最后的输出结果。

- 增加的译文结果：在待翻译短语 P 的候选译文中，支持度最高的候选译文是 W^{STOP_1} 中的高频干扰词（的、地、得、は、が、を）等，受到词长等限制不输出结果，即排在前面的译文不符合要求；在使用 W^{STOP_1} 后将这些高频干扰词去除，将次支持度的正确的候选译文项作为输出结果，提高了召回率。例如：

中文待翻译短语 P =“经典的”，$P \leftrightarrow T_g$，T_g =<“の”，“最高水準の”，…>，“の”的支持度为 4，“最高水準の”的支持度为 2，输出结果应该为“の”，但是“の”是一个 W^{STOP_1} 中的高频干

扰词，去掉这项，因此“最高水準の”作为最后的输出结果。

➢ 减少的译文结果：在待翻译短语 P 的候选译文 T_g 中，若支持度最高的译文是 W^{STOP_1} 中的高频干扰词，则将这一候选译文项去除以后，剩余的候选译文中所有项的支持度都小于阈值或受到其他限制（词长比等），因此将其原本错的译文结果去除，输出为空，即没有得到译文结果，使获取的译文结果尽量保持较高的准确率。例如：

中文待翻译短语 P =“失意的”，$P \leftrightarrow T_g$，T_g =<“チーム”，“る”，…>，“チーム”的支持度为 3，“る”的支持度为 1，输出结果应该是“チーム”，但是由于在体育领域的语料库中，“チーム”出现频率极高，因此是一个 W^{STOP_1} 中的高频干扰词，所以应该去除 而次支持度的候选译文“る”的支持度小于阈值，因此对于输入的待翻译短语“失意的”，没有对应输出的译文结果。

3.1.3 Model_with_W^{STOP_1}_W^{STOP_2} 与 Model_with_W^{STOP_1} 的比较

为了检验 Model_with_W^{STOP_1}_W^{STOP_2} 对 Model_with_W^{STOP_1} 的改进效果，分别对其在全部 40000 个短语的测试集上进行了测试，得到如下数据。

表 3 Model_with_W^{STOP_1}_W^{STOP_2} 与 Model_with_W^{STOP_1} 的比较结果

	Model_with_W^{STOP_1}	Model_with_W^{STOP_1}_W^{STOP_2}
修改后正确的结果	--	232
修改后仍错误的结果	--	39

可以看到，在模型同时使用了高频干扰词限制模块的 W^{STOP_1} 、W^{STOP_2} 后，能够修正短语译文结果中高频干扰词（如：は，を，が等）带来的影响，提高了译文准确率。下表是修改前后译文结果的实例。

表 4 使用 W^{STOP_2} 前后的译文结果对比

	奥运会 50 米自由泳金牌	大洋洲两个国家	体育预算
+ W^{STOP_1}	オリンピックで 50 メートル自由形の金メダルを	大洋州の二つ国家が	スポーツ予算は
+ W^{STOP_1} W^{STOP_2}	オリンピックで 50 メートル自由形の金メダル	大洋州の二つ国家	スポーツ予算

从表 4 中我们可以看到在没有使用 W^{STOP_2} 之前，高频干扰词的影响是确实存在的，使用了 W^{STOP_2} 之后，很好的解决了这一问题，验证了高频干扰词限制模块的有效性。

3.2 支持度限制模块的验证

为了验证支持度限制模块，并评价其对译文结果正确性的影响，随机在 40000 个待测试的中文短语中抽出 300 个短语在 Model_with_W^{STOP_1}_W^{STOP_2} 模型上进行了如下两个实验。

Test_0 为无论候选译文的支持度大小为多少，都输出支持度最大的译文作为结果。

Test_1 为当的候选译文中支持度最大值大于阈值 θ_1 时，只输出支持度最大的候选译文。

表 5 支持度限制模块的实验结果

	有译文结果的短语	译文正确的短语	译文错误的短语	准确率
Test_0	111	95	16	0.856
Test_1	58	52	6	0.895

通过实验结果可以发现，Test_1 比 Test_0 的译文正确率有所提高。这是因为，Test_0 没有考虑支持度过低时译文的准确性问题，在译文的支持度较低的情况下，待译短语在语料语料库中出现次数过低导致求交结果不够理想，从而使输出的短语译文的准确率较差，阈值的作用为排除那些最不理想的译文结果，保证了本方法的准确率，为翻译出的由多策略翻译系统的其他方法。因此对候选译文的支持度下限设定阈值是可行的，并且可以根据双语语料库 BC 及具体的译文获取任务决定阈值。

对于支持度限制模块中输出两个译文结果的情况，我们用部分短语进行了测试，发现确实有一些的次支持度的候选译文为正确译文，验证了此限制的有效性。

3.3 基于序列相交的短语译文获取方法在真实语料中的效果

序列相交模型是针对语料库的，为了检验本方法在当前双语语料库 BC 对真实语料中短语译文的翻译效果，进行了如下实验。由于本方法使用的是体育领域的双语语料，因此在新浪体育新闻中人工抽取了 100 个中文短语，在 Model_with_W^{STOP_1}_W^{STOP_2} 上进行了测试。

表 6 本方法在真实语料中的翻译效果

	有结果的短语	译文正确的短语	译文错误的短语	准确率
100 个中文短语	21	17	4	0.81

通过上表我们发现，本方法对于在真实语料中人工抽取的 100 个中文短语，有 21 个得到了译文，其中 17 个结果正确。由此可见，本方法在对双语语料库 BC 相似领域的短语译文获取方面有着一定的效果，译文结果准确率较高。但是由于当前语料库的限制，短语覆盖率较低，使本方法对真实语料中短语译文获取的召回率较低。

4 结论及下一步工作

本文提出了基于序列相交的短语译文获取方法的基本模型，及其基础上的高频干扰词限制模块、支持度限制模块，一定程度上解决了短语译文获取对词对齐、句法分析依赖的问题，是一种低代价、准确率较高的短语译文获取方法。实验表明，对于在句子级对齐的双语语料库中出现多次的短语，本方法不需要词典、词对齐、句法分析等信息就能得到较好的译文结果。由于本方法是针对于在语料库中出现多次的短语进行求交运算，因此扩大语料库、提高短语的覆盖率对提高本方法的性能有很大帮助。构造短语模板是一种能有效的提高短语覆盖率的方法，有待于继续研究；同时高频干扰词限制模块中的高频干扰词集合 W^{STOP} 的自动获取，进一步提高译文结果的准确率以及短语译文结果正确性的自动评价方法都是今后工作的重点。

参考文献

[1] Daniel Marcu, William Wong. A Phrase-based, Joint Probability Model for Statistical Machine Translation [A]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) [C]. Philadelphia, PA. USA. July 2002:133-139.

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

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

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

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

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

[7] 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[J]. 中文信息学报，2007，21(5)：91-95.

[8] 刘冬明，赵军，杨尔弘. 汉英双语语料库中名词短语的自动对应[J]. 中文信息学报，2003，17(5)：6-12.

[9] 屈刚，陈笑蓉，陆汝占. 基于有效句型的英汉双语短语对齐[J]. 计算机研究与发展，2003，40(2)：143-149

[10] 吴宏林. 面向机器翻译的汉日文本对齐研究[D]，沈阳：东北大学，2008.