doi:10.3772/j.issn.1002-0470.2013.01.005

基于对等模式的汉-英译文调序①

张春祥②* 赵铁军** 卢志茂*** 高雪瑶****

(*哈尔滨理工大学软件学院 哈尔滨 150080)
(**哈尔滨工业大学计算机科学与技术学院语言语音教育部-微软重点实验室 哈尔滨 150001)
(***哈尔滨工程大学信息与通信工程学院 哈尔滨 150001)
(****哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)

摘 要 为了提高机器翻译质量,提出了一种基于对等模式的汉-英译文调序方法:从短语翻译对中抽取汉-英语序对应关系,利用语言学特征和错误驱动学习相结合的方式获取对等模式,使用对等模式来改变汉语句法树结构,使其生成的译文符合英语语序要求。使用该方法对500个汉-英双语句对中的汉语句子进行调序的实验结果表明,词链交叉率降低了10.56%。经过调序之后,汉语句子的译文质量有所提高。

关键词 对等模式,译文调序,短语翻译对,词链交叉率,译文质量

0 引 言

译文调序是机器翻译研究中的一个关键性问题,对于提高译文输出质量和获取一致性的翻译知识具有重要的作用。目前,基于短语和句法的统计翻译模型已成为机器翻译领域的研究热点。不受限调序是一个NP难题,而N元文法语言模型不足以解决翻译中的调序问题。因此,在统计机器翻译中,引入独立的调序模型是有必要的。Tillmann在短语互译对上引入方位信息,构建了基于短语的二元文法调序模型[1]。Yamada在目标语句法树上建立统计翻译模型,从中训练出结点调序概率[2]。Nagata将两个短语的顺序关系划分为单调邻接、单调间隔、反向邻接和反向间隔四种类型,同时定义了短语之间的调序概率[3]。在IBM调序约束下,Koehn提出了基于距离惩罚的调序模型[4]。Wang设计了句法调序规则集[5]。Al-Onaizan使用词汇对齐结果和Bleu评分来度量两种语言之间的词序相似程度,定义了界外调序、界内调序和词对调序,并构建了译文调序模型[6]。Ni利用词法特征来建立调序模型,以预测短语位置概率[7]。Galley提出了一种层次化的位置模型,使用Shift _ Reduce算法来获取局部短语的语序[8]。Xiong参照词语边界信息开发了一个CKY风格的调序模型[9]。Zhang使用边界词、词性信息和短语依存关系来指导短语的调序过程[10]。Yamamoto对源语言句法树实施ITG约束,对词语调序产生了更强的限制[11]。在IST-ITG约束下,Hashimoto利用源语言句法树中的句法信息来估计目标语词序的概率[12]。本文从短语翻译对的汉语短语中抽取句法、词性、词法和词形信息作为消歧特征,依据英语短语和汉语短语之间的对译关系来获取译文调序动作,使用错误驱动的学习方法从中提取对等模式以处理汉语短语的译文调序问题。实验结果表明:采用本文所提出的方法对测试数据集中的汉语句子进行译文调序,汉-英双语句对的词链交叉率有所降低,调序后的汉语句子更符合英语语法要求。

1 汉-英语法异构

汉英语言学现象之间往往存在着较大的差异。这种差异主要表现为:在双语句对的词汇对齐结果中,存在着词链交叉现象。针对汉-英双语句对"我将在上午十点订一个预约。<->I will make an appointment at ten in the morning."而言,使用汉语句法分析器对汉语句子进行分析,利用词对齐工具对双

① 863计划(2006AA010108)和国家自然科学基金(60903082, 60975042)资助项目。
② 男,1974年生,博士,副教授,硕士生导师;研究方向:自然语言处理和机器翻译;联系人,E-mail: z6c6x6@yahoo.com.cn
(收稿日期:2012-02-16)

语句对进行词汇对齐,其结果如图 1 所示。

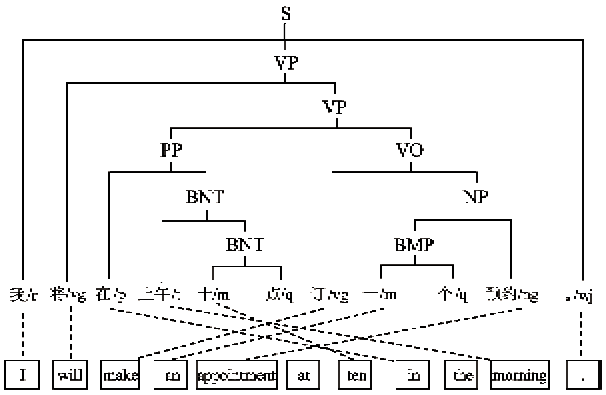

图 1 词链交叉导致汉英语序发生变化

图 1 中 S 表示句子结构;VP 为动词短语;PP 为介词短语;VO 为动宾短语;BNT 为基本时间名词短语;NP 为名词短语;BMP 为基本数量短语。从图 1 可以发现:共有 6 条词链发生交叉;在汉语句法树的第 4 层中,介词短语结点 PP 和动宾短语结点 VO 的英语译文位置正好相反;在汉语句法树的第 6 层中,词汇结点“上午/t”和基本时间名词短语结点 BNT 的英语译文位置也正好相反。词链之间的交叉是导致汉、英短语语序产生不一致的重要原因。

2 对等模式获取

在短语翻译对中,蕴含着大量的汉-英语序转换关系,而词汇对齐技术正是挖掘这种转换关系的有效手段。因此,可以利用语言学特征,从短语翻译对中抽取汉-英语序对应关系,来定义译文调序模型。本文使用对等模式来调整汉语语序。对等模式包括源模式和目标模式两部分。其中,源模式是从汉语短语中获取的;目标模式是从汉-英语序对应关系中抽取的,即译文调序动作。为便于实现对等模式的抽取,将短语翻译对进行形式化描述。

汉-英短语翻译对:Phrtype($CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$)-> e_1, e_2, …, e_n。其中,Phrtype 为汉语短语的句法标注,Phrtype($CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$)为汉语短语。$CPhrase_i$(i = 1, 2, …, m)可为词单元,仅包含汉语单词及其词性,其形式为 W_i/pos_i,W_i 表示汉语单词,pos_i 代表 W_i 的词性;$CPhrase_i$ 也可以为汉语短语。此处,e_1, e_2, …, e_n 为左部汉语短语对应的英语译文。例如:“VO[订/vg NP[BMP[一/m 个/q]预约/ng]]-> make an appointment”是一个汉-英短语翻译对。

汉语短语中蕴含的丰富的句法、词性、词法和词形信息可用于学习对等模式的源模式。为了能够获取更加抽象的源模式,需要提取 $CPhrase_i$(i = 1, 2, …, m)的核心结点的句法、词性、词法和词形特征。确定 $CPhrase_i$ 核心结点的具体过程如下:

(1)建立 $CPhrase_i$ 的句法树;

(2)后序遍历句法树,每次遇到非叶结点时,将其右子结点设置为它的核心结点,每次遇到叶结点时,将其自身设置为核心结点;

(3)遍历结束时,根结点记录 $CPhrase_i$ 的核心结点。

针对汉语短语“VO[订/vg NP[BMP[一/m 个/q]预约/ng]]”,其核心结点句法特征为 NP,核心结点词性特征为 ng,核心结点词法特征为 Object,核心结点词形特征为“预约”。

对短语翻译对 Phrtype($CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$)-> e_1, e_2, …, e_n 和 B 而言,B 为汉语短语和英语短语之间的词汇对齐结果,其译文调序动作的确定过程如下:

(1)初始化译文调序动作映射数组 A;

```
For(i=1; i<=n; i++)
A[i] = -1;
```

(2)对 $e_i \in (e_1, e_2, \cdots, e_n)$,若 e_i 在 $CPhrase_j$(j =0, 1, 2, …, m)的某个汉语单词的词典义项中出现,则 A[i] =j;若 e_i 与 $CPhrase_j$(j =0, 1, 2, …, m)的某个汉语单词 W_k 之间存在词汇对齐关系,即(i, k) ∈ B,则 A[i] =j;

(3)DecreaseConflict(A, n);

(4)Tar _ pattern =“”,输出译文调序动作;

```
For(i=1; i<=n; i++)
  IF(A[i]! = -1)
    Tar _ pattern = Tar _ pattern + A[i] + ' * '。
```

DecreaseConflict(A, n)的具体步骤如下:

```
For(i=1; i<=n; i++)
{
```

(1)寻找与 A[i]具有相同数据元素的 A[j],且 A[i]! = -1,i<j;

(2)在 A 中,从位置 i 到位置 j,寻找不相同的数据元素 A[k];

(3)如果存在不相同的数据元素 A[k],则有两种情况:

①如果 k <= (i +j)/2,则将 A 的位置为 i +2 至 j 的元素全部置为 -1,A[i] = A[k];

②如果 k > (i +j)/2,则将 A 的位置为 i +1 至

j-1的元素全部置为-1,A[j]=A[k];

}

在DecreaseConflict(A, n)中,将消除具有相同译文调序动作的片段。在这一过程中,可能会出现个别不连续的情况,例如:从位置i到位置j的所有元素,除了第k(i<k<j)个位置元素之外,所有的元素都相同。当k<=(i+j)/2时,A[k]应该移动到连续片段的左侧;当k>(i+j)/2时,A[k]应该移动到连续片段的右侧。

在短语翻译对"BNT[上午/t BNT[十/m 点/q]]->ten in the morning"中,"上午"与"morning"对齐,"十"与"ten"对齐。其中,"上午/t"为0号结点,BNT[十/m 点/q]为1号结点。初始化时,译文调序动作映射数组A[1]=-1,A[2]=-1,A[3]=-1,A[4]=-1。因为"上午"与"morning"对齐,"十"与"ten"对齐,所以,A[1]=1,A[2]=-1,A[3]=-1,A[4]=0。因此,该短语翻译对的译文调序动作为1:*+0:*。

在汉语短语中,共包含4种语言学特征:句法特征、词性特征、词法特征和词形特征。表1给出了4个基本数量短语的句法特征、词性特征、词法特征和词形特征的值。从表1中可以看出:"一家"、"这趟"、"哪种"以及"两杯"都是基本数量短语。因此,

表1 语言学特征及所覆盖的语言学现象

语言学特征	覆盖的语言学现象
句法特征	Cate=BMP 一/m 家/q,这/r 趟/q,哪/r 种/q,两/m 杯/q
词性特征	Cate=m 一,两;Cate=r 这,哪;Cate=q 家,趟,种,杯
词法特征	Head=Object 家,杯; Head=Determ 这;Head=Wh 哪
词形特征	W=一 一;W=家 家; W=这 这;W=趟 趟;W=哪 哪; W=种 种;W=两 两;W=杯 杯

当句法特征值为BMP时,可以覆盖这4个基本数量短语。词性值为m的单词包括"一"和"两";词性值为r的单词包括"这"和"哪";词性值为q的单词包括"家"、"趟"、"种"和"杯"。词法特征值为Object的单词包括"家"和"杯";词法特征值为Determ的单词包括"这";词法特征值为Wh的单词包括"哪"。每个词形特征的值仅能覆盖一个单词。从表1中可以发现,这4种语言学特征的概括能力由高到低,所覆盖的语言学单位由大到小,所覆盖的语言学现象由抽象到具体。

本文使用错误驱动学习方法从汉-英短语翻译对中获取对等模式[13]。在获取源模式时,应该以句法特征和词性特征为主,当不能对同类语言学现象进行区分时,再依次使用词法特征和词形特征。其核心思想是:首先获取抽象对等模式,抽象对等模式的源模式仅使用句法特征或词性特征,目标模式是最能概括该类语言现象的译文调序动作。对于短语翻译对Phrtype($CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$)->e_1, e_2, …, e_n和对等模式P_s->P_t而言,若Phrtype($CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$)与P_s匹配,则执行译文调序动作P_t来改变$CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$的译文生成顺序,并获取其英语译文e'_1, e'_2, …, e'_m。在e'_1, e'_2, …, e'_m与e_1, e_2, …, e_n之间,若英语单词$e'_i=e_j$,则在e'_i与e_j之间建立一条词链,从而实现两个译文片段之间的词汇对齐。对于英语译文片段"a red apple"和"an apple red"而言,存在着词链交叉,如图2所示。

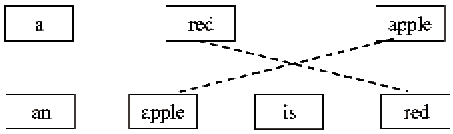

图2 英语译文片段之间的词链交叉

当e'_1, e'_2, …, e'_m和e_1, e_2, …, e_n之间存在词链交叉时,执行译文调序动作改变$CPhrase_0$, $CPhrase_1$, …, $CPhrase_m$的译文生成顺序会产生错误。此时,应该依次利用词法特征和词形特征对抽象对等模式的源模式进行特殊化处理,同时,使用调序动作生成算法从该短语翻译对中获取译文调序动作,以生成新的对等模式。反复执行这一过程,不断获取新的对等模式,直到对训练数据集中的所有短语翻译对的汉语短语都能进行正确的译文调序为止。

对以下汉-英短语翻译对,使用错误驱动的学习方法所获取的对等模式如下所示:

短语翻译对:

BNT[九月/t BNT[五/m 号/q]]->September fifth

BNT[四月/t BNT[7/m 号/q]]->April seventh

BNT[上午/t BNT[11/m 点/q]]->11:00 in the morning

对等模式:

#BNT 0:Cate=t+1:Node=BNT->0:*+1:*

#BNT 0:Cate=t+0:W=上午+1:Node=BNT->1:*+0:*

3 汉语句法短语调序

在使用对等模式“#VP 0:Node = PP + 1:Node = VO- > 1:* + 0:*”和“#BNT 0:Cate = t + 0:W = 上午 + 1:Node = BNT- > 1:* + 0:*”之后,汉语句法树的结构变化如图 3 所示。

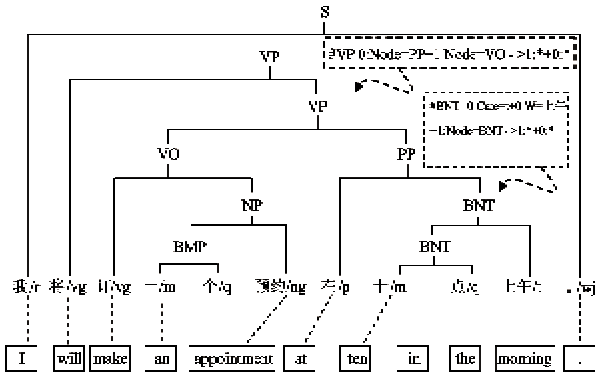

图 3 应用对等模式后的汉语句法树

从图 3 中可以发现:经过译文调序之后,双语句对之间不存在词链交叉。词链交叉数量的减少将会使上层汉语短语的译文满足英语语序的要求,可以直接对调序后的汉语句子进行直译,从而提高了译文输出质量。

4 实验及数据分析

为了分析汉-英语言学现象之间的不一致性,搜集了 44197 个汉-英双语句对。使用词对齐工具对双语句对进行词汇对齐。对汉-英双语句对进行统计分析,其结果如表 2 所示。

表 2 汉-英双语句对的性能

	数量
双语句对数目	44197
汉语单词数	353075
平均汉语句长	7.989
英语单词数	355810
平均英语句长	8.051
词链总数	246515
平均词链数	5.5776
词链交叉数	101560
平均词链交叉数	2.2979
词链交叉率	41.20%

平均汉语句长的计算公式为

$$\text{平均汉语句长} = \frac{N_1}{N_2} * 100\% \quad (1)$$

其中,N_1 为所有汉语句子所包含单词数之和,N_2 为汉语句子的数目。

平均英语句长与平均汉语句长的计算公式相似。平均词链数的计算公式为

$$\text{平均词链数} = \frac{N_3}{N_4} * 100\% \quad (2)$$

其中,N_3 为词链总数,N_4 为双语句对的数目。

平均词链交叉数的计算公式为

$$\text{平均词链交叉数} = \frac{N_5}{N_4} * 100\% \quad (3)$$

其中,N_5 为词链交叉数。

词链交叉率的计算公式为

$$\text{词链交叉率} = \frac{N_5}{N_3} * 100\% \quad (4)$$

使用汉语句法分析器对汉语句子进行分析,依据词汇对齐结果抽取汉-英短语翻译对。根据汉语短语的类型进行分类统计。非嵌套短语翻译对中的词链交叉情况如表 3 所示。

表 3 非嵌套短语翻译对中的词链交叉

	总数	词链总数	词链交叉数	词链交叉率
BAP	13677	8867	269	3.03%
BDP	276	163	5	3.07%
BMP	18764	12926	226	1.74%
BNP	38893	42059	2496	5.93%
BNS	1190	1219	123	10.09%
BNT	6502	5684	455	8.00%
BVP	18718	9708	514	5.29%

嵌套名词短语翻译对中的词链交叉情况如表 4 所示。

表 4 嵌套名词短语翻译对中的词链交叉

	总数	词链总数	词链交叉数	词链交叉率
NDE	669	818	151	18.46%
NP	22580	42029	15038	35.78%
NS	190	291	78	26.80%
NT	1946	2569	415	16.15%

嵌套动词短语翻译对中的词链交叉情况如表 5 所示。

表5 嵌套动词短语翻译对中的词链交叉

	总数	词链总数	词链交叉数	词链交叉率
VBA	1413	5437	3037	55.86%
VBEI	444	680	182	26.76%
VC	4092	8456	2326	27.51%
VJ	1916	6697	2630	39.27%
VO	45653	110168	37880	34.38%
VOO	2070	6756	2012	29.78%
VSUO	161	53	2	3.77%
VV	854	1582	392	24.78%
VP	32928	81257	34944	43.00%

其它类短语翻译对中的词链交叉情况如表6所示。

表6 其它类短语翻译对中的词链交叉

	总数	词链总数	词链交叉数	词链交叉率
PFP	5114	7088	2131	30.06%
PP	10865	12626	1472	11.66%
SS	7856	20612	6470	31.39%
AP	4451	4403	1034	23.48%
ASIDE	69	132	65	49.24%
CO	163	759	134	17.65%
DP	4	8	1	12.50%
INP	408	748	191	25.53%
MP	1097	1519	267	17.58%

由表3、表4、表5和表6可以发现:在VBA类型的短语翻译对中,词链交叉情况最严重,达到了55.86%;在ASIDE、VP、VJ、NP、VO、SS和PFP类型的短语翻译对中,词链交叉情况比较严重,都超过了30%。其原因是:这几种类型的短语,其汉语语法构成情况与英语语法构成情况存在着较大的差异。在BMP类型的短语翻译对中,词链交叉情况最小,仅有1.74%;在BAP、BDP、VSUO、BVP和BNP类型的短语翻译对中,词链交叉率不超过6%。其原因是:这几种类型的短语,其汉语语法和英语语法的定义基本一致。

本实验使用错误驱动的学习方法从短语翻译对中抽取对等模式,然后,选择源模式具有抽象概括能力的对等模式。同时,经过人工校正与筛选,选出342个对等模式。为了检验译文调序的性能,从旅游领域中抽取了500个汉-英双语句对作为测试语料,进行了以下实验:首先利用词对齐工具对这500个汉-英双语句对进行词汇对齐;然后使用汉语句法分析器对其中的汉语句子进行分析;最后应用对等模式来实现汉语句子的译文调序。分别对调序前后的双语句对的词链交叉情况进行统计分析。调序前后,测试语料的词链交叉情况如表7所示。

表7 调序前后测试语料的词链交叉

	调序前	调序后
词链总数	2641	2641
词链交叉数	917	638
词链交叉率	34.72%	24.16%

在译文调序之前,测试语料的词链交叉率为34.72%;在译文调序之后,词链交叉率仅为24.16%。使用对等模式进行译文调序,词链交叉率有了一定程度的降低,调序后的汉语句子更符合英语语法的要求。

为了进一步检验对等模式的译文调序能力,又进行了以下两组实验。在实验1中,使用基于短语的统计机器翻译系统[14]对测试语料中的汉语句子进行翻译。在实验2中,利用同一个基于短语的统计机器翻译系统对测试语料中的调序后的汉语句子进行翻译。将测试语料中的英语句子作为参考译文,使用Bleu评测方法对两组实验中的机器译文进行评价[15],其结果如表8所示。从表8中可以看出:实验2的5元Bleu评测分数和3元Bleu评测分数均高于实验1,经过译文调序之后,测试语料中汉语句子的机器译文输出质量有所提高。

表8 2组实验的译文评测分数

	Bleu5	Bleu3
实验1	0.0230974	0.0372072
实验2	0.0232820	0.0376064

5 结 论

从汉语短语中提取语言学信息作为消歧特征,从短语翻译对中获取汉-英语序对应关系;使用错误驱动的学习方法从短语翻译对中抽取对等模式;利用对等模式来改变汉语句法树的结构,使其生成的译文符合英语语序要求,是一种有效的汉-英译文调序方法。使用该方法对500个汉-英双语句对中的汉语句子进行译文调序,词链交叉率有所下降,其译文评测分数有所上升。

参考文献

[1] Tillmann C, Zhang T. A localized prediction model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Michigan, USA, 2005. 557-564

[2] Yamada K, Knight K. A syntax-based statistical translation model. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 2001. 523-530

[3] Nagata M, Saito K. A clustered global phrase reordering model for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006. 713-720

[4] Koehn P, Knight K. Feature-rich statistical translation of noun phrases. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003. 311-318

[5] Wang C, Collins M. Chinese syntactic reordering for statistical machine translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech, 2007. 737-745

[6] Al-Onaizan Y, Papineni K. Distortion models for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006. 529-536

[7] Ni Y Z. Handling phrase reorderings for machine translation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2009. 241-244

[8] Galley M. A simple and effective hierarchical phrase reordering model. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Hawaii, USA, 2008. 848-856

[9] Xiong D Y, Liu Q, Lin S X. Maximum entropy based phrase reordering model for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006. 521-528

[10] Zhang D D, Li M, Li C H. Phrase reordering model integrating syntactic knowledge for SMT. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech, 2007. 533-540

[11] Yamamoto H, Okuma H. Imposing constraints from the source tree on ITG constraints for SMT. In: Proceedings of the 2nd ACL Workshop on Syntax and Structure in Statistical Translation, Columbus, USA, 2008. 1-9

[12] Hashimoto K, Yamamoto H, Okuma H. Reordering model using syntactic information of a source tree for statistical machine translation. In: Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation, Colorado, USA, 2009. 69-77

[13] Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, 1995, 21 (4): 543-565

[14] Koehn P. Pharaoh: a beam search decoder for phrased-based statistical machine translation models. In: Proceedings of the Association of Machine Translation in the Americas, Washington, USA, 2004. 115-124

[15] Papineni K, Roukos S, Ward T. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, USA, 2002. 311-318

Reordering the translations from Chinese into English based on peer patterns

Zhang Chunxiang*, Zhao Tiejun**, Lu Zhimao***, Gao Xueyao****

(*School of Software, Harbin University of Science and Technology, Harbin 150080)

(**MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

(***College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001)

(****School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080)

Abstract

Translation reordering in machine translation was studied for improving the quality of translation output. A method for reordering the translations from Chinese into English based on peer patterns was put forward below: the correspondent relationships between Chinese word orders and English word orders were extracted from phrase translation pairs, and then, the peer patterns, obtained based on the combination of linguistics features and error-driven learning, were used to change the structures of Chinese parsing trees, making the translations accord with English word orders. The new method was used to reorder Chinese sentences of 500 Chinese-English bilingual sentence pairs in an experiment, and the results showed the crossing rate of word links degraded 10.56%, and the translation quality was increased because the Chinese sentences were reordered.

Key words: peer patterns, translation reordering, phrase translation pairs, crossing rate of word links, translation quality