从双语语料中获取翻译模板

张春祥[1],梁颖红[2],于林森[3]
ZHANG Chun-xiang[1],LIANG Ying-hong[2],YU Lin-sen[3]

1.哈尔滨理工大学 软件学院,哈尔滨 150080
2.苏州市职业大学 计算机系,江苏 苏州 215104
3.哈尔滨理工大学 计算机科学与技术学院,哈尔滨 150080
1.School of Software,Harbin University of Science and Technology,Harbin 150080,China
2.School of Computer Engineering,Vocational University of Suzhou City,Suzhou,Jiangsu 215104,China
3.College of Computer Science and Technology,Harbin University of Science and Technology,Harbin 150080,China
E-mail:z6c6x6@yahoo.com.cn

ZHANG Chun-xiang,LIANG Ying-hong,YU Lin-sen.Acquisition of translation template from bilingual corpus.Computer Engineering and Applications,2010,46(1):106-108.

Abstract: Automatic acquisition of translation templates is very important for MT system to improve its translation quality and its ability of adapting to new domain.In this paper,tree-to-string method is applied to extract translation equivalences.Error-driven learning method is used to acquire translation templates.A knowledge optimization tool is used to filter translation templates.Then these templates are applied to a transfer-based MT system,and "863" dialog corpus is used as open test corpus.The experiment shows that when new acquired and optimized templates are used,evaluation score for translation of open test corpus is improved.
Key words: translation template;translation equivalence;error-driven

摘 要:翻译模板自动获取是提高MT译文输出质量和领域适应能力的关键性因素。利用Tree-to-String方法抽取等价对,使用错误驱动的学习方法从中获取翻译模板并进行优化。将优化后的翻译模板用于一个基于转换的机器翻译系统中,同时使用"863"对话语料对其进行评测。实验结果表明:当使用自动获取并经优化的模板进行翻译时,开放测试语料的译文评测分数有一定程度的提高。
关键词:翻译模板;等价对;错误驱动
DOI:10.3778/j.issn.1002-8331.2010.01.033 **文章编号**:1002-8331(2010)01-0106-03 **文献标识码**:A **中图分类号**:TP391.2

1 引言

翻译模板是一种实现译文选择、调序的翻译知识,在统计机器翻译系统[1]和多引擎相融合的翻译系统[2]中有着广泛的应用。翻译模板是从等价对中获取的,而等价对则是从经过短语对齐的双语句对中抽取的。John证明了短语对齐最优解的搜索过程是NP难问题,同时将最优解的搜索转换为整数线性规划问题来进行求解[3]。Zhang为双语句对建立二维互信息矩阵,将互信息值相似的矩形区域视为等价对[4]。Wong采用树-串对齐(Tree-to-String)来描述源语言句法树与目标语句子之间的对应关系[5]。吴德凯提出了一种双语模型,通过统计的反向转换文法在统一的语法体系下同时对双语句对进行结构分析,分析的结果直接得到了结构对齐[6]。

采用Tree-to-String方法从汉英双语语料中抽取等价对,使用错误驱动的学习方法从中获取翻译模板。实验结果表明:自动获取的模板经过优化后,评测语料的译文输出质量有了一定程度的提高。

2 Tree-to-String对齐双语句对

对汉英双语句对(C,E)而言,等价对抽取过程如下:
(1)对C进行句法分析,获得句法树T;
(2)对C和E进行词汇对齐,获得词对齐结果A;
(3)依据A获得T中每个非叶结点对应的英语译文。
针对以下汉英双语句对,等价对的抽取过程如图1所示。
汉语句子:我们想要张靠窗户的桌子。
英语句子:We want to have a table near the window.
抽取的等价对:
VO[靠/vg 窗户/ng]->near the window
NP[VO[靠/vg 窗户/ng]的/usde 桌子/ng]->table near the

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60903082);江苏省现代企业信息化应用支撑软件工程技术研究开发中心项目(No.SX200907);黑龙江省教育厅科学技术研究项目基金(No.11541045);哈尔滨理工大学青年科学研究基金(No.2008XQJZ017);哈尔滨理工大学青年拔尖创新人才。
作者简介:张春祥(1974-),男,博士,讲师,研究领域为自然语言处理,机器翻译及机器学习;梁颖红,女,博士,副教授,研究领域为自然语言处理;于林森,男,博士,副教授,研究领域为自然语言处理。
收稿日期:2008-12-30 **修回日期**:2009-02-02

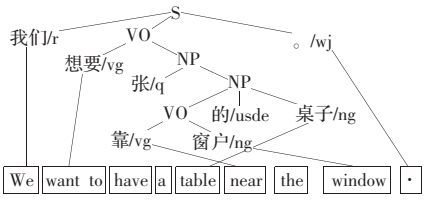

图 1　等价对的抽取过程图

window

NP[张/q NP[VO[靠/vg 窗户/ng]的/usde 桌子/ng]]->table near the window

VO[想要/vg NP[张/q NP[VO[靠/vg 窗户/ng]的/usde 桌子/ng]]]->want to have a table near the window

等价对主要包括非嵌套和嵌套两种形式：

非嵌套等价对：$Phrtype(W_0/pos_0+W_1/pos_1+\cdots+W_m/pos_m)\rightarrow(e_1,e_2,\cdots,e_n)$，其左部汉语短语由词单元 $W_i/pos_i(i=0,1,2,\cdots,m)$组成。其中 *Phrtype* 为汉语短语的句法标注；W 表示汉语单词；*pos* 代表汉语单词的词性；$e_1,e_2,\cdots,e_n$ 为对应的英语译文。

嵌套等价对：$Phrtype(PHRASE_0,PHRASE_1,\cdots,PHRASE_m)\rightarrow(e_1,e_2,\cdots,e_n)$。$PHRASE_i$ 为词单元时，仅包含汉语词及其词性，其形式为 W/pos；$PHRASE_i$ 为非嵌套短语时，由若干个词单元组成，其形式为 $Phrtype(W_0/pos_0+W_1/pos_1+\cdots+W_m/pos_m)$；$PHRASE_i$ 为嵌套短语时，可包含词单元和嵌套短语。$e_1,e_2,\cdots,e_n$ 为对应的英语译文。

3　错误驱动的翻译模板学习

翻译模板是含有变量的翻译表达式，左部是条件，右部是译文动作。翻译模板的类型与汉语短语类型完全一致。在翻译过程中，每一类模板仅作用于汉语句法分析树中具有相同类型的短语，例如：*VO* 类型的模板只用于实现 *VO* 类型短语的翻译。只有模板条件被匹配时，才会执行其译文动作。在选择翻译模板时，要求模板左部各结点的变量值被完全匹配，其中变量可以是句法特征、词性特征、词法特征和词特征。等价对左部的汉语短语用于学习翻译模板的条件，因而，在生成学习实例时，应获取等价对左部汉语短语的句法、词性、词法和词特征。

对非嵌套等价对 $Phrtype(W_0/pos_0,W_1/pos_1,\cdots,W_m/pos_m)\rightarrow(e_1,e_2,\cdots,e_n)$，通过查汉-英机器翻译词典提取汉语短语中各词单元 W_i/pos_i 的词法特征 *Head*（如：在“窗户”的词典词条中，其词法特征 *Head=Object*）。对 $(W_0/pos_0,W_1/pos_1,\cdots,W_m/pos_m)$各结点进行编号，获取该汉语短语的序号、词性、词法及词特征四元组列表$\{<i,Cate=pos_i,Head=head_i,W=W_i>,i=0,1,2,\cdots,m\}$。从 VO[靠/vg 窗户/ng]中提取的特征 4 元组列表为$\{<0,Cate=vg,Head=Svn,W=$靠$>,<1,Cate=ng,Head=Object,W=$窗户$>\}$。

对嵌套等价对 $Phrtype(PHRASE_0,PHRASE_1,\cdots,PHRASE_m)\rightarrow(e_1,e_2,\cdots,e_n)$而言，$PHRASE_i$ 可为词单元、非嵌套短语或嵌套短语。为了能获取更加抽象的翻译模板，提取 $PHRASE_i$ 核心结点的词特征及该词的词法特征用于学习过程。确定 $PHRASE_i$ 核心结点的具体过程为：(1)建立 $PHRASE_i$ 的句法树；(2)后序遍历该句法树，每次遇到非叶结点时，将其右孩子设置为它的核心结点，每次遇到叶结点时，将其自身设置为它的核心结点；(3)遍历结束时，根结点中记录了 $PHRASE_i$ 对应的核心结点。

对 $Phrtype(PHRASE_0,PHRASE_1,\cdots,PHRASE_m)$的各结点进行编号，获取该汉语短语的序号、句法、核心结点词法及核心结点词特征 4 元组列表$\{<i,Cate=Phrtype_i,Head=head_i,W=W_i>,i=0,1,2,\cdots,m\}$。从 NP[VO[靠/vg 窗户/ng]的/usde 桌子/ng]中获取的特征 4 元组列表为 $\{<0,Cate=VO,Head=Svn,W=$窗$>,<1,Cate=usde,Head=NULL,W=$的$>,<2,Cate=ng,Head=Object,W=$桌子$>\}$。

等价对中的英语片段用于学习模板的译文动作。在将汉语短语译成英语译文的过程中，通常存在着 3 种情况：选译、增译和省译。选译：即选择汉语词在词典词条中合适的义项作为译文，j:*；增译：英语译文中的某些词不能直接由汉语词翻译过来，在翻译时根据其上下文的词法和句法信息添加相应的译文，I:e；省译：某些汉语词在翻译过程中不必进行翻译。译文动作映射过程为：通过查汉英机器翻译词典确定等价对右部的英语单词，哪些属于选译情况、哪些属于增译情况。

对非嵌套等价对 $Phrtype(W_0/pos_0,W_1/pos_1,\cdots,W_m/pos_m)\rightarrow(e_1,e_2,\cdots,e_n)$，其译文动作映射过程如图 2 所示。VO[靠/vg 窗户/ng]->near the window，通过执行以上算法，其右部译文动作映射为 0:*+I:the+1:*。

(1)初始化译文动作映射数组 $A[n]$=""。
(2)对 $e_i\in(e_1,e_2,\cdots,e_n)$
①若 e_i 在 W_j 的词典义项中，且 $W_j/pos_j\in(W_0/pos_0,W_1/pos_1,\cdots,W_n/pos_m)$，则英语单词 e_i 的译文动作映射为 j:*，$A[i]=j$:*；
②若 e_i 不能在$(W_0/pos_0,W_1/pos_1,\cdots,W_m/pos_m)$中任何一个 W_j 的词典义项中找到，则该英语单词 e_i 的译文动作映射为 I:e_i，$A[i]=I$:e_i。
(3)对 $\forall i,j(A[i]=A[j]=ord$:*，具有相同结点序号的选译操作)之间的插入操作 I:进行删除，并去掉连续重复的项，使用联结符‘+’，将 A 中所有非空项连接起来输出。

图 2　非嵌套等价对的译文动作映射

对嵌套等价对 $Phrtype(PHRASE_0,PHRASE_1,\cdots,PHRASE_m)\rightarrow(e_1,e_2,\cdots,e_n)$，其译文动作映射过程如图 3 所示。NP[VO[靠/vg 窗户/ng]的/usde 桌子/ng]-> table near the window 通过执行以上算法，其右部译文动作映射为 2:*+0:*。

(1)初始化译文动作映射数组 $A[n]$=""。
(2)对 $e_i\in(e_1,e_2,\cdots,e_n)$
①若 e_i 在某个 $PHRASE_j$ 中词的词典义项中出现，则将 e_i 映射为 j:*，$A[i]=j$:*；
②若 e_i 不能在任何一个 $PHRASE_j$ 中词的词典义项中找到，则将 e_i 映射为 I:e_i，$A[i]=I$:e_i。
(3)对 $\forall i,j(A[i]=A[j]=ord$:*，具有相同序号的选译操作)之间的插入操作 I:进行删除，并去掉连续重复的项，使用联结符‘+’，将 A 中所有非空项连接起来输出。

图 3　嵌套等价对的译文动作映射

学习实例 E 与翻译模板 T 相匹配：对于学习实例 $E=SPat\rightarrow TPat$ 和翻译模板 $T=T_L\rightarrow T_R$，$SPat=\{<i,Cate=pos/Phrtype_i,Head=head_i,W=W_i>,i=0,1,2,\cdots,m\}$，$T_L=0$:$f_{00}=value_{00}+0$:$f_{01}=value_{01}+1$:$f_{11}=value_{11}+1$:$f_{12}=value_{12}+\cdots+i$:$f_{i0}=value_{i0}+i$:$f_{i1}=value_{i1}+i$:$f_{i2}=value_{i2}+\cdots+n$:$f_{n0}=value_{n0}$，若 $m=n$ 且 $f_{ij}=value_{ij}$ 出现在$<i,Cate=pos/Phrtype_i,Head=head_i,W=W_i>$中 $i=0,1,2,\cdots,m$，则 E 与 T 相匹配，E 与 T_L 相匹配。

若使用学习实例 $E=SPat\rightarrow TPat$ 的源模式 $SPat=\{<i,Cate=pos/Phrtype_i,Head=Value_i,W=W_i>,i=0,1,2,\cdots,m\}$中所有结点特征去构造模板的条件 T_L，以 T_R 作为译文动作，则获取的模板与等价对一一对应，该模板的条件 T_L 如下所示：

$T_L=0$:$Cate=pos_0/Phrtype_0+0$:$Head=head_0+0$:$W=W_0+$
1:$Cate=pos_1/Phrtype_1+1$:$Head=head_1+1$:$W=W_1+\cdots+$

$m:Cate=pos_m/Phrtype_m+m:Head=head_m+m:W=W_m$

使用错误驱动的学习方法从每一类学习实例集中获取具有一定语言现象概括能力的翻译模板。学习实例的源模式用于学习模板的条件,其目标模式用于学习模板的译文动作。学习实例的源模式中共包含 4 种特征:句法特征、词性特征、词法特征和词特征。这 4 种特征的语言现象覆盖能力呈现依次下降的趋势,因而在获取翻译模板条件时,应以句法特征和词性特征为主,当不能进行区分时,再依次使用词法特征和词特征。从某一类学习实例集 M 中获取翻译模板的算法如下:

(1)根据学习实例的抽象源模式 G 对 M 进行划分,结果记为 $M_1,M_2,\cdots,M_l$,使 M_i 中所有学习实例的抽象源模式均相同,$M_i=\{E|G(E)=G_i,E\in M\}$,$G_i$ 为 M_i 中学习实例的抽象源模式,初始化 $L_i=\Phi(i=1,2,\cdots,l)$。

(2)对 M_i 执行以下步骤:

①统计 M_i 中所有实例的目标模式,获得频度最大的目标模式 $TPat_{max}$,$L_i=L_i+\{G_i->TPat_{max}\}$,$B=\{E=SPat->TPat|E\in M_i$ 且 $TPat=TPat_{max}\}$;

②$M_i=M_i-B$,若 $M_i=\Phi$ 转(3),否则继续;

③对模板 $T\in L_i(T=T_L->T_R)$

(a)候选模板-频度列表 $CL=\Phi$,从 M_i 中获取 $M_i^T=\{E|E\in M_i$ 且 E 与 T 相匹配$\}$;

(b)从 M_i^T 所有实例的源模式中,找出不在 T_L 中出现的特征-值 $f=value$ 及 $f=value$ 的结点序号 Num 构成二元组$<Num,f=value>$,对其进行统计获取频度最大的二元组$<Num',f'=value'>$;

(c)$B=\{E|E\in M_i^T$,且 E 与 $T_L+Num':f'=value'$ 相匹配$\}$,并对 B 中所有实例的目标模式进行统计,获取频度最大的目标模式 $TPat_{max}$,其频度记为 $count$;

(d)$CL=CL+\{<T_L+Num':f'=value'->TPat_{max},count>\}$;

④从 CL 中找到 $count$ 最高的模板 $T'=T_L'->T_R'$,$L_i=L_i+\{T'\}$,$B=\{E=SPat->TPat|E\in M_i,E$ 与 T' 相匹配且 $TPat=T_R'\}$,转②;

(3)输出 $L_i(i=1,2,\cdots,l)$中的所有翻译模板。

4 实验

在 50 000 句来自旅游、新闻、交通和通用领域的汉英双语语料上,使用 Tree-to-String 方法抽取等价对。在抽取过程中所使用的词对齐工具和汉语句法分析器由哈尔滨工业大学语言语音教育部-微软重点实验室开发,其具体性能如表 1 所示。

表 1 词对齐工具和汉语句法分析器的性能分析

	精确率/(%)	召回率/(%)
词对齐工具	87	83
汉语句法分析器	78	79

共获得了 43 517 个非嵌套等价对和 95 817 个嵌套等价对。通过查汉英机器翻译词典对等价对右部进行译文动作映射,同时获得左部汉语短语的句法、词性和词法特征以生成学习实例集。对每种类型的学习实例,使用错误驱动的学习方法获得该种类型的翻译模板。从 31 类学习实例中,共获得 69 371 条翻译模板。使用基于译文评价的模板优化工具[7]对其进行优化过滤,共获得了 28 621 条翻译模板。

将优化过滤后的 28 621 条翻译模板用于一个基于转换的 MT 系统中,以检测其性能,该系统由哈尔滨工业大学语言语音教育部-微软重点实验室开发。该系统原始规则库中包含 5 092 条规则,这些翻译规则是 20 名语言工程师,经过 2 年时间编写调试出来的。为了评价获取模板的性能,使用“863”对话测试语料进行开放测试。为了比较手工获取的规则与自动获取模板的质量,共进行了两组实验。实验 1 翻译系统使用原始规则库对测试语料进行翻译;实验 2 翻译系统使用自动获取的模板对测试语料进行翻译。利用 5 元 Nist 和 Bleu 评测方法对其机器译文进行评价,其结果如表 2 所示。

从表 2 可以发现,自动获取并经优化的翻译模板在开放测试中,其各项评测分数均超过了手工书写的规则。自动获取的翻译模板其数量要超过手工书写的规则,这是因为使用机器学习方法还不能获取像语言工程师所编写的具有高度抽象概括的规则。但从获取的代价和译文质量提升方面来讲,翻译知识自动获取是可取的。这种模板自动获取方法对于在新领域中构建翻译系统的知识库而言,是极其重要的。

表 2 两组实验中开放测试语料的译文评测分数

	实验 1	实验 2
Nist5	5.752 6	5.763 3
Bleu5	0.140 7	0.148 5

5 结论

利用 Tree-to-String 方法从汉英双语句对中抽取等价对,使用错误驱动的学习方法从中获取翻译模板。同时利用基于译文评价的模板优化工具对获取的翻译模板进行优化过滤,并将其应用于基于转换的机器翻译系统中。实验结果表明:自动获取的模板,经过优化后,开放测试语料译文评测分数有一定程度的提高。在新领域中,可以自动地构建翻译系统的知识库。

参考文献:

[1] Xia Fei,McCord M.Improving a statistical MT system with automatically learned rewrite patterns[C]//Proc of the 20th International Conference on Computational Linguistics,2004:508-514.

[2] Akiba Y,Watanabe T.Using language and translation models to select the best among outputs from multiple MT systems[C]//Proc of the 19th International Conference on Computational Linguistics,2002:8-14.

[3] DeNero J,Klein D.The complexity of phrase alignment problems[C]//Proc of ACL-08:HLT,2008:25-28.

[4] Zhang Y,Vogel S,Waibel A.Integrated phrase segmentation and alignment model for statistical machine translation[C]//Proc of International Conference on Natural Language Processing and Knowledge Engineering,2003.

[5] Wong Fai,Hu Dong cheng.A flexible example annotation schema:Translation corresponding tree representation[C]//Proc of the 20th International Conference on Computational Linguistics,2004:1079-1085.

[6] Wu D.Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J].Computational Linguistics,1997,23(3):377-404.

[7] 张春祥.基于短语评价的翻译知识自动获取研究[D].哈尔滨:哈尔滨工业大学,2007.