词汇化调序模型中融合语言特性的层次短语翻译方法研究*

张 阳[1] 余正涛[2] 周 珂[1]
（1. 昆明理工大学信息科学与自动化学院 昆明 650500）（2. 昆明理工大学智能信息处理重点实验室 昆明 650500）

摘 要 针对越南语语言特性，提出在词汇化调序模型中融合语言差异特性的汉语-越南语的统计机器翻译方法。该方法首先分析汉语与越南语语法不同，提取越南语在定语位置、状语位置及修饰词词语顺序上与汉语的差异，然后形式化定义这些差异规则，以对数线性模型的形式融入进词汇化调序模型中。在训练过程，通过融合语言差异特性的词汇化调序模型对符合特性的规则进行权重调优，从而在解码过程中指导候选翻译的选择。实验结果表明，在词汇化模型里融合语言特性的汉语-越南语的层次短语机器翻译模型比基准系统提高了 0.6 ~ 2.1 个 BLUE 值。

关键词 统计机器翻译；词汇化调序模型；汉语；越南语；语言特性
中图分类号 TP391 **DOI**：10. 3969/j. issn. 1672-9722. 2017. 12. 014

A Lexicalized Reordering Model of Integrating with Language Features for Hierarchical Phrase-based Translation

ZHANG Yang[1] YU Zhengtao[2] ZHOU Ke[1]
（1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500）
（2. Key Laboratory of intelligent information processing, Kunming University of Science and Technology, Kunming 650500）

Abstract According to the language characteristics of Vietnamese, this paper proposed a new lexicalized reordering model where language features were integrated for Chinese-Vietnamese statistical machine translation. Firstly, the grammar differences between Chinese and Vietnamese were analyzed, and the sequence differences in attribute, adverbial modifier and adjuncts were extracted. Secondly, the extracted difference rulers were formally defined and be integrated in the lexicalized reordering model via the log-linear model. In the training processing, the proposed model would optimize the weight for these rules that conform to the linguistic features Finally, it would guide the translation selection in the decoding. The experiment had verified that our reordering model achieved a 0.6-2.1 BLEU point improvements for Chinese-to-Vietnamese translation over a baseline hierarchical phrase-based system.

Key Words statistical machine translation, lexicalized reordering model, chinese, vietnamese, language features
Class Number TP391

1 引言

越南是中国西南的重要邻国之一，研究汉语-越南语统计机器翻译对汉越双语理解、舆情分析、文化交流等具有重要的支撑作用。目前，汉越统计机器翻译研究的工作主要集中在汉越双语语料库建设、汉越词对齐研究等方面，翻译方法研究尚处于起步阶段。汉语-越南语词对齐方面，莫媛媛等[1]将越南语定语后置特征融入对数线性模型，显著提高了词对齐的准确率。越南语翻译方面，美国空军在越战期间资助开发了第一个英-越翻译系统 LOGOS I，该系统中仅包含 130 个词，只是初步具有翻译功能；周云[2]使用基于规则的句法转换方法开发了汉-越机器翻译系统 Han Viet 0.1，该系统只能转换局部子树而不能转换全局子树，其规则也未能覆盖全部语言现象；Phuoc Tran 和 Dien Dinh[3]提出了

* **收稿日期**：2017 年 6 月 8 日，**修回日期**：2017 年 7 月 28 日
作者简介：张阳，男，硕士研究生，研究方向：自然语言处理，机器翻译。余正涛，男，教授，博士生导师，研究方向：机器翻译，自然语言处理。周珂，男，硕士研究生，研究方向：自然语言处理，机器翻译。

基于汉-越意义关系的SVBUT模型和PVBUT模型,主要解决未知词的翻译问题;Google翻译中也集成了汉-越翻译系统,但其翻译质量还远不能满足需求。以上汉-越机器翻译方法并没有考虑汉语、越南语的语言特征对翻译效果的影响,而不同语言的差异性对机器翻译系统的性能具有不可忽略的影响,因此如何将汉语和越南语的语言差异性融合到机器翻译系统中是一个值得研究的问题。

从语法差异性上来说,越南语和汉语在语法[4~8]上既有相同点又有不同点,相同点是两种语言主干成分的语序一致,均为SVO(主-谓-宾),例如,汉语"我听音乐",越语为"Tôi(我)nghe(听) nhạc(音乐)";不同点在于,越南语有些定语后置特性。举例如下,汉语句子"她是美丽的女孩",而越南语句子则是Cô là(她是)cô gái(女孩)đẹp(美丽的)。通过形式化的定义这些规则,将其融合到基于层次短语的翻译方法中的词汇化模型[9~11]中,以此来探究汉语和越南语的语言特性对翻译系统性能的影响。

2　词汇化模型的层次短语翻译方法

2.1　层次短语翻译模型

层次短语模型[12]是基于同步上下文无关文法的翻译模型,其层次短语的形式如下表示:

$$X \rightarrow <r, a, \sim> \tag{1}$$

其中,X 是层次短语翻译模型中使用的唯一一个代表短语的非终结符,γ 和 α 都是由终结符和非终结符构成的串。~代表 γ 和 α 中非终结符的一对一的对齐信息。层次短语通过四个步骤从双语例句中学习得到:首先获取双语句对的词对齐信息,接着从中提取初始短语对,然后抽取带有非终结符的层次短语对,最后计算层次短语对的翻译概率。基于获取的层次短语模型,翻译模型用CYK算法对待翻译句子进行分析的同时生成候选翻译,然后从候选译文中选择得分高的译文作为最终译文。假设待翻译句子为 f_1^L,f_1^L 和匹配的层次短语集合为 Z_1^K,那么最佳的译文 e_1^I 可由下面的式(2)得到:

$$e_1^I = \arg\max_{e_1^I, z_1^K} \{\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^L, z_1^K)\} \tag{2}$$

其中,$h_i(e_1^I, f_1^L, z_1^K)$ 是特征函数,λ_i 是权重。常用特征包括:词汇翻译和层次短语转换的概率模型、译文长度模型、目标语言模型等。

2.2　调序关系

层次短语的调序[13]都是通过规则完成,只需定义规则的调序就能直接获得翻译的调序。仅考虑变量与前后词语的位置关系,因此只需分析泛化规则。主要是由于短语规则本身就是完全词汇化,其调序并没有歧义性。因此我们设定两种调序方向分别为单调(Monotone)及交换(Swap)。单调调序是指词语在越南语端和汉语端的相对位置并没有改变,而交换调序则表示相对位置发生改变,如图所示。图1使用规则(3)进行翻译的情况。图中变量 X_1 与前面的短语对(我,Tôi)在汉语端和越南语端的位置关系都是一致的,因此是单调调序。然而汉语端 X_1 位于"放进"之前,越南语端 X_1 的位置被调整到了"trong"后面,因此这里的调序是交换调序。规则:

$$X \rightarrow <\text{我}\ X_1\ \text{放进}\ X_2,\ \text{Tôi trong}\ X_2\ X_1> \tag{3}$$

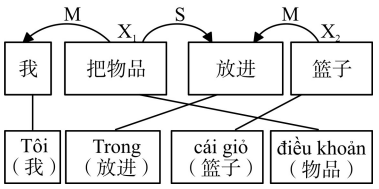

图1　使用规则(3)进行翻译的调序情况

形式化上,给定一个规则及其词语对齐关系 $a = \{(i, j)\}$,根据变量与源端邻接词语的位置关系,定义四种调序关系。变量的对齐表示为 (i^x, j^x);源端词语的对齐为 $a^w = \{(i^w, j^w)\}$;源端邻接单词用W表示。考虑两种情况,前一个词 (-1),后一个词 $(+1)$;词序方向用O表示。根据W、O的取值情况,定义四种调序关系:

$$(W=-1, O=M): i^w < i^x \ and \ \forall (i^w, j^w) \in a^w : j^w < j^x;$$
$$(W=-1, O=S): i^w < i^x \ and \ \exists (i^w, j^w) \in a^w : j^w > j^x;$$
$$(W=+1, O=M): i^w < i^x \ and \ \forall (i^w, j^w) \in a^w : j^w > j^x;$$
$$(W=+1, O=S): i^w < i^x \ and \ \exists (i^w, j^w) \in a^w : j^w < j^x.$$

2.3　词汇化调序模型

通过变量所泛化短语的边界词估计上述四种调序关系的概率。一个变量在汉语端和越南语端有四个边界词分别是:汉语端左边界词(clb);汉语端右边界词(crb);越南语端左边界词(vlb);越南语端右边界词(vrb)。如图1中变量 X_1 四个边界词分别是:clb=把;crb=物品;vlb=vrb= điều khoản(物品),按照如下方式估计变量 X_1 与前一个词"我"的调序的概率:

$$P(O|W=-1, clb=\text{把},\ crb=\text{篮子},\ vlb=\text{điều khoản},\ vrb=\text{điều khoản}) \tag{4}$$

给定变量的边界词以及变量与邻接词语的调

序方向。

我们估计参数 $P(O|W, c\,\mathrm{lb}, crb, v\,\mathrm{lb}, vrb)$。这组参数同时考虑汉语端和越南语端的边界词语。到此仅说明了规则上的调序情况,完整的词汇化调序模型[]是定义在推导上的。具体定义如下:

与左边邻接词语的调序模型:

$$P_{-1}=\prod_{r\in D}\prod_{x\in r}P(O|W=-1, c\,\mathrm{lb}, crb, v\,\mathrm{lb}, vrb) \quad (5)$$

与右边邻接词语的调序模型:

$$P_{+1}=\prod_{r\in D}\prod_{x\in r}P(O|W=+1, c\,\mathrm{lb}, crb, v\,\mathrm{lb}, vrb) \quad (6)$$

其中 D 表示推导,r 表示规则,x 表示 r 中的变量。

概率估计使用词对齐的语料来估计参数。当一个规则被抽取的时候,同时记录各种调序实例,以此来估计参数。如图 1 所示,当规则(3)被抽取的时候,记录如下的实例:

O=M, W=−1, clb= 把, crb= 物品, vlb= điều khoản, vrb= điều khoản

O=S, W=+ 1, clb= 把, crb= 物品, vlb= điều khoản, vrb= điều khoản

O=S, W=−1, clb=篮子, crb=篮子, vlb= cái giỏ, vrb= cái giỏ

当抽取完实例,就可以用这些实例来估计上述的条件概率。

3 融合越南语特性的词汇化调序模型

3.1 越南语特定分析

越语和汉语修饰语(定语、状语)与中心语的排列顺序的差异[4~8]就是这两种语言最明显的不同点。对于修饰语的位置问题,汉语里修饰名词性成分的定语一律排在名词之前,而越语里定语一般位于名词之后。汉语里描写性多层递加定语的排列顺序是:1)谓语短语;2)动词(短语)\介词短语;3)形容词短语及其他描写性短语;4)不用“的”的形容词和描写性名词。越语中描写性多层定语结构顺序与汉语呈镜像关系,汉语中描写性定语的顺序是:1–2–3–4 中心语,越南语的顺序是:中心语 4–3–2–1。表 1 为越—汉词语排列结构特点示例。

表 1 越–汉词语结构特点

汉语词序	越南语词序
她是我见过的最美丽的女孩	Cô là(她是)cô gái(女孩)đẹp(美丽的)nhất(最)mà tôi từng thấy(我见过的)
孩子们很喜欢那种上面有字的白色的心状糖果	Trẻ con(孩子们)rất(很)thích(喜欢)loại(那种)kẹo(心状糖果)màu trắng(白色)ben trên có chữ(上面有字)

针对越南语的语言特点,将越南语的语法特征总结为以下几点:

1)修饰动词的副词置后,即修饰动词的副词位于动词的后面;

2)修饰形容词的副词置后,即修饰形容词的副词位于形容词的后面;

3)修饰名词的形容词置后,即修饰名词(或者名词性短语)的形容词位于名词(或者名词性短语)的后面;

4)多个形容词修饰同一名词,形容词逆序置后,即多个形容词共同修饰一个名词(或者名词性短语),形容词的顺序逆序(相对于汉语中的顺序)排列,并且位于所修饰的名词后面。

基于上述规律,我们定义三种调序规则:

1)动词+副词;

2)形容词+副词;

3)名词(名词性短语)+形容词。

3.2 融合特性的词汇化模型

我们使用北大中文词性标注集[14]。北大中文词性标注集中名词、形容词、动词、副词等等,其标志如下表 2 所示。

表 2 汉语和越南语中的词性标志对应关系

词性	汉语标志
名词	n(普通名词)、nr(人名)、ns(地名)、nt(机构团体)、nz(其他专名)、r(代词)
形容词	a(形容词)
副词	d(副词)
动词	v(动词)

我们使用这些词性对应关系,分析在 2.3 节抽取出来的调序实例,对符合我们定义越南语特性的实例进行概率调优。

举例如下:

汉语:她是美丽的女孩。

越南语:Cô là(她是)cô gái(女孩)đẹp(美丽的)。

通过层次短语翻译模型,抽取到这样的规则:$X\rightarrow$<她是X_1, Cô là X_1>,由于层次短语过度泛化,致使 X_1 缺失了调序信息。针对这个问题,我们引入词汇化调序模型,基于 2.3 节抽取出来的实例:O=M, W=−1, clb=美丽的, crb=女孩, vlb= cô gái, vrb= đẹp,通过北大中文词性标注可以识别出这个实例是符合我们定义的第三种调序规则。

根据北大词性标注工具,可以将 3.1 节的三种调序形式化定义为

1)V + D(动词+副词),记为 VD

2) A + D(形容词+副词),记为AD

3) N + A (名词(名词性短语)+形容词),记为NA

融合语言特性的左邻接调序模型 P_{-1} 为:

$$P_{-1}=P_{-1(原)}\times\exp(\lambda_{VD}\cdot VD)\times\exp(\lambda_{AD}AD)\times\exp(\lambda_{NA}NA) \quad (7)$$

λ 表示调序规则所对应的权重。

4 实验

4.1 实验数据

由于现在缺乏权威的汉语-越南语双语句对平行语料库,实验所使用的汉越平行语料均是来自实验室内部,均是从互联网爬取汉-越篇章对齐语料,然后通过计算机句对齐和人工校对的方法获取的。目前,收集的汉越平行句对约有30000余条。实验数据的详细信息如下表3所示。

表3 实验中所用语料信息

		中文	越南语
训练集	句数	25000	
	词数	275659	275659
开发集	句数	4000	
	词数	47962	47962
语言模型训练集	句数	2000	
	词数	22763	22763
测试集	句数	1000	
	词数	11142	12169

4.2 实验设计及结果分析

实验以Niutrans开源系统中的短语模型系统[15]作为基准原型,使用本实验室自主开发的工具对越南语进行词法分析,并且使用GIZA++得到汉语-越南语的词对齐结果。三元语言模型和四元语言模型的训练借助于Niutrans机器翻译平台。使用BLEU[16]评测标准作为评测工具。实验中我们设计三组对比实验来检验融合语言特性的词汇化调序模型对汉语-越南语机器翻译的影响,分别是:汉语-越南语层次短语的机器翻译方法HPB(Hierarchical Phrase-Based SMT),融合词汇化调序模型的汉语-越南语层次短语的机器翻译方法HPB+MS(Hierarchical Phrase-Based SMT + MS),融合语言特性的词汇化调序模型的汉语-越南语层次短语的机器翻译方法HPB+MS(Grammar)(Hierarchical Phrase-Based SMT + MS(Grammar)),语言模型选取二元、三元和四元。式(7)中的 λ 符合语言特性条件的 $\lambda=1$,反之 $\lambda=0$。实验结果如表4所示(实验结果评估采用三元语言模型)。

表4 汉语-越南语机器翻译系统对比实验

方法	NIST	BLEU(%)
2-HPB	4.1009	32.86
3-HPB	4.1557	24.76
4-HPB	4.1757	18.36
2-HPB + MS	4.3125	34.88
3-HPB + MS	4.3788	25.38
4-HPB + MS	4.3902	20.50
2-HPB + MS(Grammar)	4.4107	35.74
3-HPB + MS(Grammar)	4.4559	26.87
4-HPB + MS(Grammar)	4.4867	21.69

从表4可知,仅使用基于汉语-越南语层次短语的机器翻译方法,翻译结果是比较差的,原因可能是不能覆盖比较全面的语言现象,未能抽取出调序信息,不能指导解码过程;仅用词汇化调序的汉语-越南语层次短语的机器翻译方法,翻译效果有所改善,但是不能解决语言特性对翻译效果的影响;将语言特性融入进词汇化调序模型中的汉语-越南语短语的机器翻译时,翻译效果得到了明显改善。实验结果表明,融合语言特性的词汇化模型的汉语-越南语层次短语统计机器翻译方法能够有效的提高翻译的准确率。

5 结语

针对越南语典型的语言后置的特点,提出一种融合语言特性的词汇化模型的统计机器翻译方法。该方法在汉语-越南语的层次短语模型的基础上,通过训练过程,将语言特性融入词汇化调序模型,对符合语言差异的实例进行权重调优,从而在解码过程中指导候选翻译的选择。实验证明了提出方法的有效性,融合语言特性的词汇化调序模型能够提高翻译的准确率。进一步研究主要集中在基于句法的汉越双语翻译模型中融合语言特性。

参考文献

[1] Yuanyuan Mo, Jianyi Guo, Zhengtao Yu, etc. A bilingual word alignment algorithm of Vietnamese-Chinese based on feature constraint[J]. International Journal of Machine Learning & Cybernetics, 2015(6):538-543.

[2] 周云. 汉语越南语机器翻译实验系统[D]. 洛阳:解放军外国语学院,2006.
ZHOU Yun. The Chinese to Vietnamese Machine Translation experimental system[D]. Luoyang: PLA University of Foreign Language, 2006.

[3] Phuoc Tran, Dien Dinh, Linh Tran. Resolving Named Enti-

(下转第2427页)

[8] 严建峰,李伟华,刘明. 基于混合蚁群算法的MAS任务分配[J]. 计算机应用研究,2009,26(1): 68-70.
YAN Jianfeng, LI Weihua, LIU Min. Task allocation for MAS based on hybrid ant colony algorithm[J]. Application Research of Computers,2009,26(1): 68-70.

[9] 胡敏,杨伟民,林春晓. 基于公共对象请求代理结构的负载测试技术研究[J]. 上海理工大学学报,2007,29(2):189-194.
HU Min, YANG Weimin, LIN Chunxiao. Research on the load testing system based on CORBA and its implementation[J]. University of Shanghai for Science and Technology,2007,29(2):189-194.

[10] 刘治国,孙连宝,王祎. 海战场电磁环境仿真系统及数学模型研究[J]. 舰船科学技术,2010,32(4):130-133.
LIU Zhiguo, SUN Lianbao, WANG Yi. Research on the mathematical models and complex EM environment simulation system of the sea battlefield[J]. Ship Science and Technol0gy,2010,32(4):130-133.

[11] 王鹏,王莉,高斌. 综合电子信息系统生成能力评估方法[J]. 指挥控制与仿真,2014,36(5):67-71.
WANG Peng, WANG Li, GAO Bin. Survivability Evaluation Method of C4ISR[J]. Command Control & Simulation,2014,36(5):67-71.

(上接第 2392 页)

ty Unknown Word in Chinese-Vietnamese Machine Translation[J]. Advances in Intelligent Systems & Computing, 2014,(245):273-284.

[4] 闫丹辉,毕玉德. 基于规则的越南语命名实体识别研究[J]. 中文信息学报,2014,28(5):198-205.
YAN Danhui, BI Yude. Rule-based Recognition of Vitnamese Named Entities[J]. Journal of Chinese Information Processing,2014,28(5):198-205.

[5] 武氏春蓉. 略论汉语对越南语的影响[J]. 济南大学学报(社会科学版),2001,11(5):56-57.
WU ShiChunRong. A Brief Comment On the Influence of Chinese on Vietnamese[J]. Journal of Jinan University (Social Science Edition),2001,11(5):56-57.

[6] 武氏河. 越南语与汉语的句法语序比较[J]. 云南师范大学学报对外汉语教学与研究版,2005,3(6):65-68.
WU Shihe. A Comparison of the Syntactic Order of Vietnamese and Chinese[J]. Journal of Yunan Normal University(Teaching and Research on Chinese As A Foreign Language Edition),2005,3(6):65-68.

[7] 何氏红凤. 汉语和越南语"定语"的对比分析[D]. 武汉:华中师范大学,2006.
HE ShiHongFeng. Contrastive analysis on "attributes" in Chinese and Vietnamese[D]. Wuhan:Huazhong University of Science and Technology,2006.

[8] 范青钏. 汉越状语语序比较分析[J]. 云南师范大学学报对外汉语教学与研究版,2007,5(6):79-85.
FAN Qingchuan. Contrastive analysis on "adverbial" in Chinese and Vietnamese[J]. Journal of Yunan Normal University(Teaching and Research on Chinese As A Foreign Language Edition),2007,5(6):79-85.

[9] Christoph Tillmann. A unigram orientation model for statistical machine translation[C]//Proceedings of HLT-NAACL 2004: Short Papers. Association for Computational Linguistics,2004:101-104.

[10] Koehn P, Axelrod A, Mayne A B, et al. 2005. Edinburgh system description for the 2005 NIST MT evaluation[C]//Proceedings of Machine Translation Evaluation Workshop.

[11] Jinsong Su, Yang Liu, Yajuan Lu, etc.Learning Lexicalized Reordering Models from Reordering Graphs[C]// Proceeding of Meeting of the Association for Computational Linguistics. 2010:12-16.

[12] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 263-270.

[13] 肖欣延,刘洋,刘群,等. 面向层次短语翻译的词汇化调序方法研究[J]. 中文信息学报,2012,26(1):37-41.
XIAO Xinyan, LIU Yang, LIU Qun, et al. Lexical Reordering for Hierarchical Phrase-based Translation [J]. Journal of Chinese Information Processing,2012,26(1): 37-41.

[14] 俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报,2002,16(5):49-64.
YU Shiwen, DUAN Huiming, ZHU Xuefeng, et al. The Basic Processing of Contemporary Chinese Corpus at Peking University Specification[J]. Journal of Chinese Information Processing,2002,16(5):49-64.

[15] Tong Xiao, Hao zhang, Qiang Li, et al. The niutrans machine translation system for cwmt2011[C]//Proceeding of the 6th China Workshop on Machine Translation, Xiamen, China, August 2011. CWMT.

[16] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics,2002:311-318.