软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(12):3088−3100 [doi: 10.3724/SP.J.1001.2012.04207]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

层次短语翻译中基于 Markov 随机场的层次切分模型*

刘乐茂[1,2+], 赵铁军[1,2], 曹海龙[1,2], 朱聪慧[1,2], 张春越[1,2]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](语言语音教育部-微软重点实验室(哈尔滨工业大学),黑龙江 哈尔滨 150001)

Hierarchical Partition Model Based on Markov Random Fields for Hierarchical Phrase-Based Machine Translation

LIU Le-Mao[1,2+], ZHAO Tie-Jun[1,2], CAO Hai-Long[1,2], ZHU Cong-Hui[1,2], ZHANG Chun-Yue[1,2]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](MOE-MS Key Laboratory of Natural Language Processing and Speech (Harbin Institute of Technology), Harbin 150001, China)

+ Corresponding author: E-mail: lmliu@mtlab.hit.edu.cn

Liu LM, Zhao TJ, Cao HL, Zhu CH, Zhang CY. Hierarchical partition model based on Markov random fields for hierarchical phrase-based machine translation. *Journal of Software*, 2012,23(12):3088−3100 (in Chinese). http://www.jos.org.cn/1000-9825/4207.htm

Abstract: The partition ambiguity of translation derivations is an important problem suffered by the statistical machine translation, and it is much more important in a hierarchical phrase-based machine translation. In the paper, a hierarchical partition model is proposed to address the problem. The study applies markov random fields to construct the model, and integrate it into the hierarchical translation model to automatically select the more reasonable partition. In the NIST Chinese-English translation tasks, the optimization of the model is very efficient, and it improves the translation performance for hierarchical phrase-based translation on NIST05, NIST06 and NIST08 test sets.

Key words: hierarchical phrase translation; partition model; graphical model; Markov random fields; dependency tree

摘 要: 翻译推导的切分歧义是统计机器翻译面临的一个很重要的问题,而在层次短语机器翻译中,其尤为突出.提出了一个层次切分模型来处理推导的切分歧义性.采用 Markov 随机场构建模型,然后将其融入层次短语翻译模型,以便自动选择更合理的切分.在 NIST 中英翻译的任务中,该模型的训练效率高,通过 NIST05,NIST06 和 NIST08 这 3 个测试集上的翻译效果表明,该模型提高了层次短语翻译的性能.

关键词: 层次短语翻译;切分模型;图模型;Markov 随机场;依存树

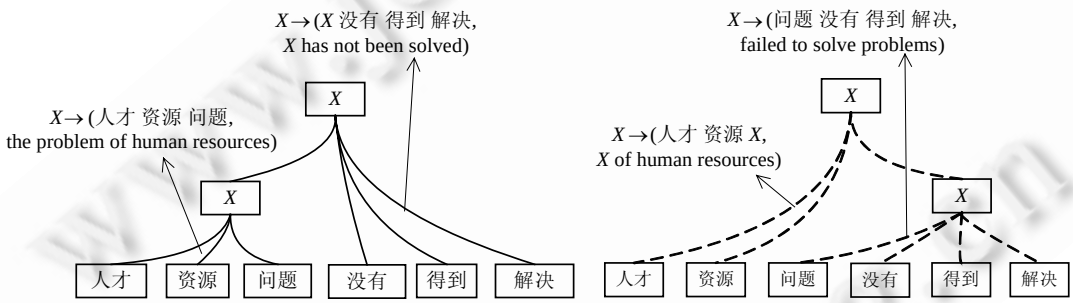
中图法分类号: TP391 **文献标识码**: A

迄今为止,统计机器翻译取得了很大的发展,涌现出许多机器翻译模型.其中,主流模型有基于短语的模型[1,2]和基于句法的模型[3].层次短语翻译模型是一种特殊的句法模型,不同于纯句法翻译模型[3−5],它不带有任

* 基金项目: 国家自然科学基金(60736014, 61173073, 61100093); 国家高技术研究发展计划(863)(2011AA01A207)
收稿时间: 2011-07-14; 修改时间: 2011-11-02, 2012-01-16; 定稿时间: 2012-03-19

何语言学上的句法意义,从形式上讲,它依赖于同步上下文无关文法,是一种形式句法模型[5,6],因此不会像纯句法模型那样遭遇双语对齐约束与语言学句法约束的冲突.同时,层次短语模型还兼顾了短语和纯句法模型的优点.一方面,它的翻译单元是层次化的短语,同短语一样,层次化短语能够有效地捕捉到局部的固定搭配和调序,从而保证翻译在局部上有较好的流利度;另一方面,与纯句法模型相同,层次短语翻译规则允许非终结符的存在,这使得它能够处理非连续短语之间长距离的调序.

在统计机器翻译中,推导是翻译模型中的隐变量,一个目标翻译对应于许多翻译推导.要获得给定源语言的目标翻译,需要对与它相对应的所有推导求和.实际上,绝大多数的翻译系统的解码器都采用最大推导的方式来近似翻译概率(见公式(4)).这样,翻译推导的歧义问题就成为了制约翻译性能的最重要原因之一.与基于短语的翻译模型相比,层次短语模型的翻译规则含有非终结符,导致其翻译规则的数量大大增加.这样就使得翻译推导歧义性问题更为严重:源语言端,对于给定的句子片段而言,存在众多不同的切分方式;在目标语端,源语言端相同的切分含有许多不同的翻译.文献[5]的翻译模型可以理解为一个全局的、隐式的翻译推导的消歧模型,但是这种消歧模型的能力有限.因此,本文的想法是显式建立一个子模型,然后通过它约束翻译推导中的翻译规则,进而实现对翻译推导消歧.本文考虑的是翻译推导在源语言端的歧义问题,称为切分歧义问题.图 1 展示了一个翻译例子,其中:实线表示的推导经过源语言和参考译文的双语句法分析而来,它对应的翻译是参考译文;虚线的推导是层次短语翻译模型的输出,这两个推导在源语言端的切分方式不同.



实线对应的是参考译文: The problem of human resources has not been solved.
虚线对应的是层次短语系统输出: Failed to solve problems of human resources.

Fig.1 Two different derivations for a source sentence

图 1 一个源语言句子的两个不同推导

对于层次短语翻译而言,一个源语言 $f_0^n = f_0 f_1 \cdots f_{n-1}$ 的推导 d 在源语言端的映射对应于源语言上的一个层次切分(hierarchical partition,简称 HP),它是 f_0^n 的一个覆盖集合,并且满足任何两个覆盖是:要么其中一个包含另外一个,要么互不相交.为叙述方便,在本文中,覆盖和切分表达相同的含义.图 1 中的实线对应的推导的层次切分是{[0,3],[0,6]}.本文通过对这样候选翻译推导对应的层次切分 HP 进行建模 $p(HP)$,文中称其为层次切分模型,以便达到对层次翻译推导消歧的目的.延续上下文无关文法的思想,在做一些独立性假设的前提下,本文将 $p(HP)$进行局部分解:

$$p(HP) = \prod_{span \in HP} p(span) \tag{1}$$

这样做的一个好处是,与全局特征(层次切分模型)相比,局部特征更有利于翻译过程的解码搜索.至此,关键问题是如何定义其中的因子(局部特征)$p(HP)$,文中称它为切分模型,本文尝试使用概率图模型中的 Markov 随机场(MRF)[7]对其进行解释.对源语言句子中的覆盖进行概率建模,也即对其中的词序列建模,会遭遇一个难题:由于每个词对应于概率模型中的一个随机变量,随着词序列长度的增加,概率图模型的推理会更加复杂,比如边缘化、归一化运算,进而影响翻译的速度.为了克服这个困难,本文的设想是,从词序列中选择几个具有代表性意义的词进行建模.考虑到依存结构反映的是词与词在语义上的依赖关系,同时受文献[8]中心词驱动思想的影响,

本文选择的是词序列中的中心词.正如前所述,由于层次短语翻译规则不具有语言学上句法信息,它的翻译规则的覆盖可能并不遵循语言学上的句法边界,这样可能导致一个覆盖的中心词在这个覆盖之外.为此,本文首先从依存树中寻找出这个词序列对应的头节点序列(定义见第 1 节),通过对这个头节点序列建模来近似表达 $p(HP)$.在翻译过程中,它通过评价依存头节点序列的概率,约束翻译规则歧义切分,以便让解码器输出更好的候选翻译.另外,它的训练和解码高效,几乎不会影响原有翻译系统的效率.据我们所知,本文首次使用图模型的方法构建服务于机器翻译的子模型.

本文第 1 节简单介绍基线系统-层次短语翻译模型.第 2 节、第 3 节中分别给出切分模型的定义和推理、训练.第 4 节介绍如何把模型加入翻译解码器.第 5 节给出本文的实验.第 6 节回顾相关工作.最后是本文的结论.

1 层次短语机器翻译

层次短语机器翻译基于概率同步上下文无关文法(PSCFG).从形式上讲,PSCFG 是一个五元组$\langle N,T_s,T_t,R,w\rangle$,其中,$N$是非终结符集合,$T_s$,$T_t$分别表示源语言和目标语言端的终结符集合,$R$是文法中的产生式规则集合,$w$是$R$上的一个非负的实值权重函数.$R$ 中的每个产生式规则 r 形如 $X\rightarrow\langle\alpha,\gamma,\sim\rangle$,其中,$X\in N$ 是一个非终结符,表示规则的左部;$\langle\alpha,\gamma,\sim\rangle$表示规则的右部,$\alpha\in(N\cup T_s)^*$表示规则源语言端,$\gamma\in(N\cup T_t)^*$为规则的目标语言端,$\sim$表示$\alpha$和$\gamma$中非终结符的对应关系.在搭建翻译系统中,层次规则集 R 可以从双语对齐语料上自动进行抽取.在基于层次短语翻译的 Hiero[5,6]系统中,规则的非终结符不代表任何的语言学信息,仅仅表示一个需要扩展的占位符.除了开始的非终结符(S)外,它含有一个非终结符.除了这些规则外,Chiang 还定义了两种 glue 规则:$S\rightarrow\langle S\ X,S\ X\rangle$和 $S\rightarrow\langle X,X\rangle$.

在翻译过程中,给定一个句子 f 以及 PSCFG 文法,翻译的目的就是在候选翻译构成的空间$\Delta(f)$中寻找一个翻译作为目标输出 $\hat{e}$,使得

$$\hat{e}=\arg\max_{e\in\Delta(f)}p(e\mid f) \tag{2}$$

其中,$\Delta(f)$由 PSCFG 文法进行解码产生.$p(e|f)$的定义依赖于产生式规则中的权重函数,理论上

$$p(e\mid f)=\sum_{d\in D(f,e)}p(d)=\sum_{d\in D(f,e)}\prod_{r\in d}w(r) \tag{3}$$

其中,$D(f,e)$表示能够产生(f,e)的同步推导集合,d 表示其中的一个推导,它是由一些产生式规则 r 构成的树.在解码过程中,由于对所有 $d\in D(f,e)$求和的代价过高,实际上,使用最大推导近似 $p(e|f)$

$$\hat{e}=e\left(\arg\max_{d\in D(f)}\prod_{r\in d}w(r)\right) \tag{4}$$

其中,$D(f)$表示产生源语言 f 的所有推导集合,$e(d)$表示推导 d 对应的目标翻译.在层次短语机器翻译系统中,基于最大熵的框架,$w(r)$的定义由一系列特征函数按照如下方式表示:

$$w(r)=\prod_i\phi_i^{\lambda_i}(r) \tag{5}$$

其中,ϕ_i为规则的特征函数,λ_i为其对应的权重函数,它们可以通过 MERT[9]在一组开发集上确定.Hiero 定义了如下特征函数[1,5]:

- 两个方向上的层次规则翻译概率 $p(\alpha|\gamma)$,$p(\gamma|\alpha)$;
- 两个方向上的词汇化概率 $p_{lex}(\alpha|\gamma)$,$p_{lex}(\gamma|\alpha)$;
- 层次短语规则的惩罚 P_Penalty;
- glue 规则的惩罚 G_Penalty;
- 词惩罚 W_Penalty;
- 语言模型 Lm.

层次短语模型的翻译过程可以看作利用 PSCFG 的产生式规则实现源语言到目标语言的转录(transduction),Hiero 系统采用 CYK 风格的柱搜索实现解码器.

2 基于 MRF 的切分模型定义

本文接下来的内容是构建公式(1)中的切分模型.正如前面所叙述的,理论上,切分模型应该建模一个覆盖中所有的词,但这样会引起模型的训练和翻译解码的开销过大.为了克服这个问题,本文借用依存句法分析来简化切分模型.实际上,在机器翻译中应用依存句法结构的思路屡见不鲜.文献[10]提出了一种基于路径的翻译模型,将翻译过程解释为源语言和目标语的依存树中基于路径的片段之间的转换.不同于他们的模型,文献[11]考虑的是目标语言端的依存树结构,建立了一个源语言的串到目标语言端依存树的翻译模型.依存句法结构在统计机器翻译中取得的巨大成功,是本文考虑使用依存句法的一个动机.不同于上述提出的方法,本文采用依存句法树不是产生含有句法信息的翻译规则,而是构建翻译模型中的层次切分模型,以便减小层次翻译推导在源语言端的切分歧义性.本文中使用的仅仅是源语言一端的依存结构,没有利用目标语言端依存结构信息.目标语言端的句法信息可以通过结构化语言模型的方式作用到翻译结果上,进而提高翻译的质量.在后续的工作中,我们也将考虑如何将目标语言端的依存结构纳入到模型中.由于文中切分模型的定义基于依存结构,我们先简要叙述依存结构.

设 $f = f_0^n$ 是一个源语言句子,$T(f)$是 f 的依存句法树.$T(f)$也可以表示成序列 $d_0d_1\ldots d_{n-1}$,其中,$-1\leqslant d_i\leqslant n-1$,$0\leqslant i\leqslant n-1$.如果 $d_i=j$,它表示 f_i 依赖于 f_j.特别地,当 f_i 是 $T(f)$的根节点,$d_i=-1$.在图 2 中表示了一棵依存树,有 $d_0=1$,$d_4=-1$.设 r 是 f 上的一个可用的翻译规则,并且它在 f 上的覆盖为$[a,b]$,其中,$0\leqslant a<b\leqslant n$.以下本文考虑$[a,b]$在 $T(f)$中的分布情况,粗略地说,可以分为以下两种:$T(f)$中有一棵子树恰好覆盖$[a,b]$,此即,$[a,b]$在 $T(f)$有一个头节点;任何一棵子树都不能覆盖$[a,b]$,也即,$[a,b]$在 $T(f)$没有头节点.在图 2 中,r_1 在源语言端的覆盖为[0,2],它有一个依存头节点"资源/NN";r_2 的覆盖为[2,4],它没有头节点,因为其中 2 个词的依存头节点均为"得到/VV",但是它在覆盖[2,4]之外.上述的分类方式与文献[12,13]相似.通过统计发现,规则的覆盖在源语言的依存树中没有头节点的情形发生的频率极高.本文认为其中可能原因是:一方面,规则产生所依赖的双语词对齐的不准确性;另一方面,依存句法分析的结果存在错误.下面给出统一的严格定义.

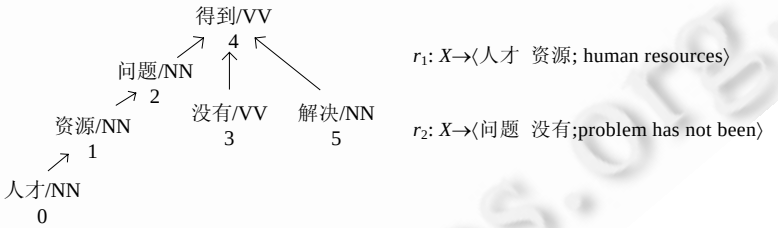


Fig.2 A dependency tree for a source sentence and its available translation rules

图 2 一个源语言的依存结构树及其可用的翻译规则

形式上说,覆盖$[a,b)$在 $T(f)$的头节点序列是一个有序的正整数集合$\{i_1,i_2,\ldots,i_k\}$,其中,$a\leqslant i_j<b$,$1\leqslant j\leqslant k$,并且满足下述条件:

- 集合中所有的元素互不相同;
- 任何 $a\leqslant u<b$,存在一个 $j(1\leqslant j\leqslant k)$,使得 $i_j=d_u$;同时,对任何 $j(1\leqslant j\leqslant k)$,存在某个 $u(a\leqslant u<b)$,使得 $d_u=i_j$.

通过定义可以发现,$\{i_1,i_2,\ldots,i_k\}$为 $T(f)$中一系列子树的根,这些子树满足:互不相交,且它们的并恰好覆盖$[a,b]$.如图 2 所示,r_1 的覆盖的头节点序列为{1};r_2 的为{2,3}.

为了定义公式(1)中的层次切分模型,只需要定义其中的因子-切分模型.正如前所述,理论上,切分模型应该定义在覆盖中所包含的词序列之上.然而,随着覆盖中词序列长度增加,概率模型的边缘化,归一化运算的复杂度会越来越大,这就需要尽量减少模型中的随机变量数目.由于依存头节点序列可以抓住覆盖中词序列的一部分语义信息,因此本文将覆盖的切分模型定义在覆盖的依存头节点序列之上.现在,本文从概率模型的角度出发给出公式(1)中切分模型 $p(span)$的一般定义:如果一个覆盖在 $T(f)$上的头节点序列为$\{h_1,h_2,\ldots,h_s\}$,这个覆盖的切

分模型解释为全概率 $p(h_1,h_2,\ldots,h_s)$,此即 $p(span)=p(h_1,h_2,\ldots,h_s)$;同时,规则 r 的切分模型定义为它在源语言端覆盖的切分模型.为了减轻参数估计的负担,一个自然的想法就是假设头节点之间满足条件独立性,即

$$p(h_1,h_2,\ldots,h_s)=p(h_1)\times p(h_2)\times\ldots\times p(h_s) \tag{6}$$

然而,这个模型具有很强的局限性——头节点之间的条件独立性.比如,若前一个头节点为动词,那么紧接着的头节点仍然为动词的可能性较小,它为名词的可能性更大.为了对不同依存头节点之间的相互依赖关系进行建模,本文借助更为一般的概率模型——图模型构建切分模型.

概率图模型是一个由顶点集合和边集合构成的图,其中每个顶点表示一个随机变量或者随机变量集合,每条边表示随机变量之间的依赖关系[7].通过引入图,概率图模型可以清晰地表达出联合概率分布的可分解性以及随机变量之间的条件独立性等等.依据图中是否含有向边进行分类,图模型可以分为有向图模型(贝叶斯网络)和无向图模型(马尔可夫随机场,Markov random fields,简称 MRF).同有向图模型相比,无向图模型更容易表达随机变量之间的软依赖关系.因为在某些情况下,人们并不关心这个变量有向依赖于那个变量,人们只关心这两个变量之间是否存在依赖关系.概率图模型中,顶点的定义是抽象的,在计算语言学中,它们可以具体为词、词性和句法结构标记等等.在当今的研究中,MRF 的一个特殊例子 CRF(conditional random fields)在自然语言处理领域有着广泛的应用:文献[14]使用 CRF 构造一个序列标注模型,在词性标注方面,实验结果表明,这个模型的性能优异;文献[15]构造了一个基于 CRF 的句法分析模型,他们的模型与其他一些性能很好的判别式句法分析模型相当,甚至更为优越.

理论上说,本文可以建立表达能力更强的、含有环的图模型.在参数优化过程中,可以借助于近似推理的方法,比如变分法或者 Loopy Belief Propagation 方法[7],但是这会增加参数估计过程中的代价.本文中只考虑链式的图模型,其中最大的团是图中的边,这样可以有效地控制推理的复杂度.下面使用 MRF 来定义切分模型:

$$p(h_1,h_2,\ldots,h_s)=\frac{\exp\left(\sum_{i=1}^{s}\sum_{k}\lambda_k\times f_k(h_i)+\sum_{i=1}^{s-1}\sum_{k'}\mu_{k'}\times g_{k'}(h_i,h_{i+1})\right)}{Z(s)} \tag{7}$$

$$Z(s)=\sum_{h_1,h_2,\ldots,h_s}\exp\left(\sum_{i=1}^{s}\sum_{k}\lambda_k\times f_k(h_i)+\sum_{i=1}^{s-1}\sum_{k'}\mu_{k'}\times g_{k'}(h_i,h_{i+1})\right) \tag{8}$$

其中,对每个 $k,k',f_k,g_{k'}$均表示特征,λ_k 和$\mu_{k'}$为它们对应的函数;s 为头节点序列的长度,$Z(s)$为归一化因子(partition function).

严格来说,规则 r 的切分模型的定义不仅与它覆盖的头节点序列有关,而且还与规则本身有关,甚至与依存树有关.即规则的切分模型应该定义成如下的条件概率:

$$p(h_1,h_2,\ldots,h_s\mid r,T(f))=\frac{\exp\left(\sum_{i=1}^{s}\sum_{k}\lambda_k\times f_k(h_i,r,T(f))+\sum_{i=1}^{s-1}\sum_{k'}\mu_{k'}\times g_{k'}(h_i,h_{i+1},r,T(f))\right)}{Z(s,r,T(f))} \tag{9}$$

$$Z(s,r,T(f))=\sum_{h_1,h_2,\ldots,h_s}\exp\left(\sum_{i=1}^{s}\sum_{k}\lambda_k\times f_k(h_i,r,T(f))+\sum_{i=1}^{s-1}\sum_{k'}\mu_{k'}\times g_{k'}(h_i,h_{i+1},r,T(f))\right) \tag{10}$$

但是,构建如公式(9)的切分模型会面临两个问题:首先,在训练中,由于 $Z(s,r,T(f))$同时与结构化的依存树 $T(f)$和层次短语规则 r 有关,特别是后者的规模达数千万级(在本文的实验中,训练语料中可以抽取出近三千万条翻译规则),这样,计算所有 $Z(s,r,T(f))$所需要的时间复杂度就远远大于千万级别**;另外一方面,更为重要的是,在解码过程中,解码器对层次规则 r 的响应频率非常高,解码速度直接与规则的响应速度有关,因此 $Z(s,r,T(f))$的计算速度就会影响解码的速度.为了简化模型的复杂度以便提高模型估计的效率,本文使用公式(7)中的联合概

** 实际上,根据第 3.2 节中的公式(20)所述,公式(8)中计算 1 次 $Z(s,r,T(f))$的复杂度为 $O(s\times|features|\times|h|^2)$,其中,$s$ 为头节点序列的长度,$|features|$为特征数,$|h|$为头节点种类数(也即图模型中节点的状态数).所以在训练中,对于千万级别的规则 r,计算所有的 $Z(s,r,T(f))$需要 $O(10^8\times s\times|features|\times|h|^2)$;而如果采用公式(7)中的与规则无关的模型,则计算复杂度仅为 $O(s\times|features|\times|h|^2)$.

率 $p(h_1,h_2,\ldots,h_s)$取代上面的条件概率.这样的做法诚然会削弱切分模型的表达能力,比如,如果规则 r_1 与 r_2 的源语言端不同,而它们在源语言端的覆盖相同(图 1 中覆盖[0,6]的两个规则 $X\rightarrow\langle X$ 没有 得到 解决,X has not been solved$\rangle$和 $X\rightarrow\langle$人才 资源 X,X of human resources$\rangle$),那么它们对应的头节点序列就相同.于是,它们按照公式(7)获得了相同的切分模型概率得分,因此,切分模型对这种情况视作没有区分度.但是,这并不会极大地降低层次切分模型的表达能力,比如,从公式(1)看,层次切分模型依然能区分这两个推导:这两个规则中的非终结符的覆盖不同,所以图中两个推导的层次切分模型依然不同.

3 推理和训练

极大似然估计是概率模型参数估计中的最重要方法之一,本节采用极大似然法训练模型中的参数λ_k,μ_j.对每个 $i(i=1,2,\ldots,n)$, $H_i=\{h_1^i,h_2^i,\ldots,h_{s(i)}^i\}$ 是一个长度为 $s(i)$的头节点序列,那么根据公式(7)、公式(8),可以得出这组样本的 log 似然函数:

$$\ell(H;\lambda,\mu)=\sum_{i=1}^{n}\log p(H_i)=\sum_{i=1}^{n}\left(\sum_{j=1}^{s(i)}\sum_{k}\lambda_k\times f_k(h_j^i)+\sum_{j=1}^{s(i)-1}\sum_{k'}\mu_{k'}\times g_{k'}(h_j^i,h_{j+1}^i)-\log Z(s(i))\right) \tag{11}$$

为了抑制参数的过拟合,本文使用高斯先验对 log 似然函数进行正则.为此,需要最小化如下目标函数:

$$\mathcal{R}(H;\lambda,\mu)=-\ell(H;\lambda,\mu)+\left(\sum_{k}\frac{\lambda_k^2}{2\sigma^2}+\sum_{k'}\frac{\mu_{k'}^2}{2\sigma^2}\right) \tag{12}$$

3.1 切分模型的推理

目标函数(12)关于参数λ,μ是凸的、光滑的,具有全局最小值,可以使用梯度的方法进行优化,其梯度为

$$\nabla_{\lambda}\mathcal{R}(H;\lambda,\mu)=-\sum_{i=1}^{n}\left(\sum_{j=1}^{s(i)}f(h_j^i)-\frac{1}{Z(s(i))}\times\frac{\partial Z(s(i))}{\partial\lambda}\right)+\frac{\lambda}{\sigma} \tag{13}$$

$$\nabla_{\mu}\mathcal{R}(H;\lambda,\mu)=-\sum_{i=1}^{n}\left(\sum_{j=1}^{s(i)-1}g(h_j^i,h_{j+1}^i)-\frac{1}{Z(s(i))}\times\frac{\partial Z(s(i))}{\partial\mu}\right)+\frac{\mu}{\sigma} \tag{14}$$

其中,$\sigma(\sigma>0)$是正则因子,f 和 g 分别是由分量 f_k 和 $g_{k'}$所组成的特征向量函数.优化的关键是计算 $Z(s)$、偏导 $\partial Z(s)/\partial\lambda$及$\partial Z(s)/\partial\mu$.在图模型中,对这些量的推理一般采用消息传递的方法,也即 CRF 和 HMM 中的前向-后向算法,其主要思想是使用动态规划算法求解[7].分别递归定义如下函数:

$$\alpha(i,h)=\begin{cases}\exp\left(\sum_{k}\lambda_k\times f_k(h)\right), & \text{if } i=1\\ \sum_{h'}\alpha(i-1,h')\exp\left(\sum_{k}\lambda_k\times f_k(h)+\sum_{k'}\mu_{k'}\times g_{k'}(h',h)\right), & \text{else } i>1\end{cases} \tag{15}$$

及

$$\beta(j,h)=\begin{cases}\exp\left(\sum_{k}\lambda_k\times f_k(h)\right), & \text{if } j=1\\ \sum_{h'}\beta(j-1,h')\exp\left(\sum_{k}\lambda_k\times f_k(h)+\sum_{k'}\mu_{k'}\times g_{k'}(h,h')\right), & \text{else } j>1\end{cases} \tag{16}$$

其中,$\alpha(i,h)$和$\beta(j,h)$的物理含义可以理解成:按照图 3(a)中的方式,一条信息沿着链自左向右从 h_{i-1} 传递到 h_i,在 $h_i=h$ 处,该信息的量值为$\alpha(i,h)$;按照图 3(b)中的方式,一条信息自右向左从 h_{j-1} 传递到 h_j,在 $h_j=h$ 处,该信息的量值为$\beta(j,h)$.公式(15)和公式(16)的表达形式十分相似,仅有的差别是公式(15)中的 $g_{k'}(h',h)$和公式(16)中的 $g_{k'}(h,h')$.从消息传递的角度看,它们的差别反映的是图 3 中消息传递时的 2 种不同方向.则有如下公式成立:

$$Z(s)=\sum_{h}\alpha(s,h)=\sum_{h}\beta(s,h) \tag{17}$$

$$\frac{\partial Z(s)}{\partial \lambda_k}=\sum_{h}\sum_{i=1}^{s}\alpha(i,h)\times\beta(i,h)\times f_k(h)\times\exp\left(-\sum_{k}\lambda_k\times f_k(h)\right) \tag{18}$$

$$\frac{\partial Z(s)}{\partial \mu_{k'}}=\sum_{h,h'}\sum_{i=1}^{s}\alpha(i,h)\times\beta(i+1,h')\times g_{k'}(h,h')\times\exp\left(-\sum_{k'}\mu_{k'}\times g_{k'}(h,h')\right) \tag{19}$$

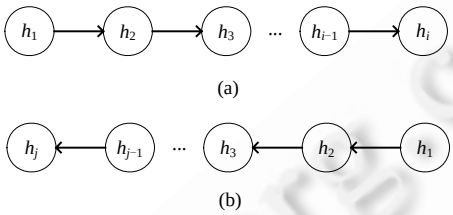


Fig.3 Message propagation

图 3 消息传递

至此,根据公式(17)~公式(19),可以有效地计算出 $Z(s)$,$\partial Z(s)/\partial\lambda$及$\partial Z(s)/\partial\mu$.

3.2 切分模型的训练

算法 1 给出了切分模型训练的伪代码(实验中使用 C++实现了该算法),其中调用了一个伪牛顿的优化工具——LBFGS[16].算法 1 中的代码第 2 行~第 16 行是一次完整的迭代过程;在当前的权重向量(λ,μ)下,第 2 行~第 9 行记录了在一次迭代过程中 $Z(s)$,$\partial Z(s)/\partial\lambda$及$\partial Z(s)/\partial\mu$的值;这样,在对每个样本计算 log 似然函数及其梯度(第 10 行、第 11 行)时可以使用这些已经计算过的量,从而可以节省大量的计算时间.特别强调的是,如果使用 CRF 构建形如公式(6)中的切分模型,计算 $Z(s,r,T(f))$及其梯度值(第 10 行~第 12 行)需要对每个样本执行一次第 2 行~第 9 行,这将会耗费大量的计算量.对于本文中所采用的 MRF 全概率模型来说,算法 1 在剔除 LBFGS 优化外,每次迭代的时间复杂度为

$$O(|features|\times|h|^2\times \mathrm{MAX}LEN+n\times|features|\times \mathrm{MAX}LEN) \tag{20}$$

其中,$|features|$表示特征的个数,即(λ,μ)的维数;$|h|$为头节点种类数(即图模型中节点的状态数),本文中表示的是词性标注集合中元素的个数;MAX*LEN* 为训练样本中头节点序列的最大长度;n 为训练样本的个数.因此,只要适当选择$|h|$,训练算法的时间复杂度与特征数和样本数是线性关系.

算法 1. 训练算法.

输入:n 个训练样本 $H_1,H_2,\ldots,H_n$;正整数 MAX*LEN*,使得对任何 i,H_i 头节点序列的长度至多为 MAX*LEN*;迭代停止阈值ρ;正则系数σ;

输出:权重(λ,μ).

1. 初始化权重向量(λ,μ)=(0,0);
2. **For each** s(1≤s≤**MAX*LEN***) //s 为头节点序列的长度
3. **For each** h //每个头节点
4. 根据递归公式(15)和公式(16)计算$\alpha(i,h)$,$\beta(s,h)$;
5. **For each** λ_k //特征权重
6. 根据公式(18)计算$\partial Z(s)/\partial\lambda_k$;
7. **For each** $\mu_{k'}$ //特征权重
8. 根据公式(19)计算$\partial Z(s)/\partial\mu_{k'}$;
9. 根据公式(17)计算 $Z(s)$;
10. 根据公式(12)计算目标函数值$\mathcal{H}$; //遍历每个样本 H_i
11. 根据公式(13)和公式(14)计算目标函数关于变量的梯度$\nabla_\lambda\mathcal{H}$,$\nabla_\mu\mathcal{H}$; //遍历每个样本 H_i

12. **LBFGS** 优化:LBFGS($\lambda,\mu,\mathcal{H},\nabla_\lambda\mathcal{H},\nabla_\mu\mathcal{H}$),返回新的权重($\lambda',\mu'$);
13. **If** ($\|(\lambda-\lambda',\mu-\mu')\|<\rho$)
14. **Then return** (λ',μ');
15. **else**
16. (λ,μ)=(λ',μ');
17. **Goto** 2.

正如公式(1)所述,本文训练切分模型是为构建层次切分模型 $p(HP)$,而层次切分的定义依赖于双语训练语料的推导,因此从理论上说,切分模型的训练例子需要从双语训练语料的推导中获取.但是获取双语推导的代价过高,因为双语句法分析的时间复杂度为 $O(|f|^3|e|^3)$[17].为此,本文使用其他方法来获取切分模型的训练例子.与文献[18−20]的方法相似,本文也从规则抽取过程中获得训练实例.注意,本文构造的是全概率模型,所以,不需要像他们训练分类器那样为每种类别构造出一些训练例子.本文的训练例子作为层次短语翻译规则抽取过程的副产品,它们的获取过程是:在规则抽取前,先要获得训练集源语言端的依存句法树;每次抽取出一个层次规则时,记录下该规则在源语言的覆盖,并计算它在依存树中的头节点序列,这个头节点序列就是一个例子.这样,随着规则抽取的完成,就获得了训练例子集合.

4 融入切分模型的解码器

总体来说,本文中的模型是以一个特征的方式融入到层次短语翻译模型中;更具体地说,是在公式(4)中额外增加一个特征函数(在翻译模型中,同其他特征(翻译概率等)一样,它对应的权重也是经 MERT 优化而来).本文提出的模型与规则在源语言端的覆盖有关,因此不能像其他的特征如翻译概率和词汇化翻译概率一样,在翻译模型训练时就可以将它统计出来,它需要在解码过程中动态地计算.在解码器每次搜索到一个可用的层次规则 r 时,根据源语言端的依存结构树,确定 r 对应的头节点序列$\{h_1,h_2,\ldots,h_s\}$,然后通过已训练出的切分模型,根据公式(7)和公式(8)计算出 $p(h_1,h_2,\ldots,h_s)$的值,并把这个值加入公式(4).正如前面提到,$p(h_1,h_2,\ldots,h_s)$的计算时间直接影响解码的速度.为了提高计算速度,与训练中的做法类似,只需要对每个长度为 s 的头节点序列,据公式(17)计算 $Z(s)$的值,然后将结果保存起来.待计算 $p(h_1,h_2,\ldots,h_s)$时,先通过查表得出 $Z(s)$的值,接着在载入的模型中查出所需的特征权重的值并计算出概率的值,这样就可以极大地加速解码.

5 实 验

5.1 实验设置

本实验中的翻译任务是 NIST 中英翻译.在实验过程中,本文使用的训练语料是 FBIS,开发集是 NIST02,测试集是 NIST05,NIST06,NIST08,具体的数据分布见表 1.在 Gigaword 语料中的新华部分(共 181M 英文单词)的单语语料上,我们使用 SRILM[21]工具训练了一个 4 元语言模型,其中,平滑算法采用修正的 Kneser-Ney 方法[22].为了训练出层次短语的翻译模型,需要获得训练语料的词对齐.为此,本文先使用 GIZA++[23]获得训练语料的两个方向的词对齐,然后使用 grow-diag-final-and 启发式规则获得两端的多对多的词对齐信息.为了评价翻译的效果,本文使用大小写敏感的 4-gram BLEU[24]和 NIST[25]两种评价度量,它们基于 n-gram 评价粒度,是国际上公认的机器翻译评价准则,本文使用集成了这两种度量的评价工具 mteval-v13a.pl(http://www.statmt.org/).

Table 1 Distribution of resource data

表 1 语料数据分布

语料名称	句对数	中文/英文词数
FBIS	239 335	6 844 069/897 383 4
NIST02	878	23 196/26 519
NIST05	1 082	30 257/31 950
NIST06	1 663	38 721/47 332
NIST08	1 357	32 712/41 104

在实验中,我们使用的基线系统是课题组内部实现的层次短语机器翻译系统,它与 Hiero 相似,对应的特征如引言中提到的.在解码过程中,基线系统的设置如下:

- *rule_limit*=10.载入翻译模型时,对源语言端相同目标语端不同的层次短语规则,解码器最多载入 top-10 个不同的规则;
- *hypo_size*=200.每个存储部分翻译假设的栈的大小设置为 200;
- *g_KBEST*=100.在使用 MERT 优化模型的参数时,解码器最多只输出 100 个候选翻译;
- *span_limit*=10.

在进行 CKY+解码扩展时,对于 *span*<10 的情况,解码器可以使用层次规则进行解码扩展;当 *span*≥10 时,解码器只能使用粘贴规则进行扩展.为了展示基线系统的性能,本文选择著名的开源机器翻译系统 Moses[2]作为参照.采用相同的数据,它们的 BLEU 得分见表 2.

Table 2 BLEU of Baseline vs. Moses

表 2 基线系统与 Moses 的 BLEU 对比

Systems	NIST05	NIST06	NIST08
Baseline	24.74	24.52	17.87
Moses	24.87	24.00	18.04

5.2 训练切分模型

为了训练本文提出的切分模型,首先使用 Stanford Parser[26]获得源语言端的完全句法结构,然后使用文献[27]的中心规则,将完全句法树转化为依存结构(在开发集和测试集上,也使用相同的方式获得依存结构树).在层次短语翻译规则获取的过程中,同时可以得到切分模型的训练例子.本次实验共有例子约 35M,实验中使用词性作为特征,不同依存头节点词性序列的部分分布情况见表 3.

从表 3 中可以看到:尽管含有一个头节点的比重更高,含有多个头节点的情形仍占有相当的比例.另外,某些多头节点的比重比一个头节点的更高,比如{NN PU}和{P}比重对比,这些证据可以从某种程度上表明本文对依存头节点序列进行建模的可行性.在训练切分模型时,本文设定正则系数为 15(相对每个训练例子而言).训练模型高效,算法 1 每次迭代需要约 2 分,迭代 20 次后收敛.

Table 3 Percent of partial different head node sequences

表 3 部分头节点序列的百分比

头节点序列	百分比(%)	头节点序列	百分比(%)
VV	16.05	NR	1.00
NN	13.03	PU	0.86
NN PU	2.95	NN PU NN	0.81
VV PU	2.44	VA	0.80
VV NN	1.92	AD	0.77
NN PU VV	1.73	VV VV	0.76
NN VV	1.67	VV PU VV	0.72
VC	1.60	PU NN	0.70
PU VV	1.53	VE	0.63
P	1.42	NN NN	0.60

5.3 翻译对比结果及分析

为了评价切分模型对翻译起到的效果,我们先在开发集 NIST02 上输出 100best,接着使用 MERT[28]训练出基线系统(baseline)中各模型的参数;然后,在基线系统中增加一个层次切分模型(基于 MRF 的得分),同样的配置下训练出另外一个系统,取名为 Plus Hierarchical Partition System(PHPS).尽管多加入了一个概率子模型,Baseline 和 PHPS 的解码速度相当.

表 4 给出了两个模型的权重,从中可以看出,层次切分模型与翻译的总得分是正相关的,同时,这也从某种程度上表明本文构造模型的合理性.

Table 4 Feature weights for two translation systems

表 4 两个翻译系统的特征权重

Systems	$p(\alpha\|\gamma)$	$p(\gamma\|\alpha)$	$p_{lex}(\alpha\|\gamma)$	$p_{lex}(\gamma\|\alpha)$	P_Penalty	G_Penalty	W_Penalty	Lm	MRF
Baseline	0.0676	0.0899	0.0557	–0.1883	0.1694	0.0133	–0.2818	0.1335	—
PHPS	0.0904	0.0630	0.0190	–0.1534	0.0610	0.1280	–0.1905	0.1033	0.1910

我们使用 3 个测试集 NIST05,NIST06,NIST08,同时使用两个翻译自动评价指标,Baseline 和 PHPS 效果对比见表 5 和表 6.

Table 5 Comparison BLEU results of 3 different systems on 3 NIST Chinese-English test sets

表 5 在 3 个 NIST 中英测试集上 3 个系统的对比结果——BLUE

Systems	NIST05	NIST06	NIST08
Moses	24.87	24.00	18.04
Baseline	24.74	24.52	17.87
PHPS	25.24	25.23	18.46

Table 6 Comparison NIST results of 3 different systems on 3 NIST Chinese-English test sets

表 6 在 3 个 NIST 中英测试集上 3 个系统的对比结果——NIST

Systems	NIST05	NIST06	NIST08
Moses	7.8688	7.5505	6.0516
Baseline	7.8705	7.6237	6.0393
PHPS	7.9151	7.7422	6.1974

从 BLEU 度量来看,在 3 个测试集上,PHPS 比 Baseline 至少提高了 0.5 个 BLEU,并且提升的最大幅度可达 0.7 个 BLEU;同时,PHPS 的 BLEU 也比 Moses 的更高.从 NIST 评价度量来看,PHPS 也同样在 3 个测试集上获得了更好的性能.因此,从整体上可以得出结论,本文提出的基于 MRF 的切分模型确实可以提高层次短语翻译的性能.我们从系统的输出中找出了几个样本,两个系统的输出的目标翻译如下:

- 样本 1

  源语言:其中 一 次 会议 的 地点 在 巴黎 , 另 一 次 会议 的 地点 尚 未 决定 .

  Baseline: A meeting in Paris , another meeting venue has not yet decided .

  PHPS: One meeting place in Paris and another venue of the meeting has not yet been decided .

  参考译文 1: One meeting will be held in Paris and the other has yet to be determined.

  参考译文 2: One of the meetings would be in Paris, and the other was yet to be determined.

- 样本 2

  源语言:不过 他 认为 , 经过 美军 长期 训练 后 , 伊拉克 部队 的 训练 会 获得 成果 .

  Baseline: However , he said , after the US long-term training Iraqi military training will be given .

  PHPS: But he believed that after US military training for a long time , the Iraqi military training will be given results .

  参考译文 1: But he believes that after being trained for a long time by the US army, Iraqi forces will achieve good training results.

  参考译文 2: But he believed the training would be fruitful after the Iraqi forces finish a long period of training by the US military.

通过比较两个系统输出结果对应的层次切分我们发现,在 2 个样本上的切分存在众多不同:比如,Baseline 翻译样本 1 时有一个切分为“的 地点 在 巴黎”,而 PHPS 的切分为“其中 一 次 会议 的 地点”;Baseline 翻译样本 2 时有一个切分为“认为 , 经过 美军”,PHPS 的切分为“不过 他 认为 ,”.这个事实表明,本文的模型确实可以通过选择更加合理的切分方式,从而获得更好的翻译结果.

6 相关工作

在基于形式文法翻译模型中融入语言学知识,是提高翻译性能的有效手段之一.文献[13]从源语言端的句法树中提取了一些成分特征,其中一些特征鼓励符合句法边界的规则,另外一些特征惩罚不符合句法的那些规则.相似地,文献[28]提出更多的语言学特征,既有源语言端的,又有目标端的.文献[29,30]提出了更为丰富的特征,不同于上述使用 MERT 优化特征权重,他们使用在线的 MIRA 算法.与本文的方法不同,上面提出的一些特征都是启发式的,因此,这些特征对于翻译规则之间的区分度不太敏感.当翻译模型中融入大量的这种特征时,系统性能的提成更加明显.文献[18,31]在形式文法的翻译模型中通过融入上下文的语境信息,建立了一个基于判别式的子模型.与本文的模型不同,他们的模型侧重于翻译规则中非终结符覆盖的边界词信息,而不关心覆盖内部的词的信息.实际上,规则覆盖内部的词的信息也同样重要.另外,本文使用 MRF 构建翻译中的子模型,同他们的基于判别式的分类器不同,本文不需要为每个类都寻找出一部分的训练样本.

依存结构反映的是语义上词与词的依赖关系,这个语义约束可以很方便地作用到翻译模型中.文献[11]将目标语言的依存结构以依存语言模型的方式作用于目标翻译;文献[32]将依存结构树的信息以启发式特征的方式加入翻译模型中;文献[33]在层次短语翻译中通过源语言端的依存结构构建了一个长距离的调序模型.而不是使用语言学上的依存结构,文献[34]根据双语词对齐的信息诱导出一个形式的依存结构,从而通过它约束解码过程产生的层次树.与上述方法不同,本文利用依存结构的目的是从规则覆盖的词序列中提出一部分语义更相近的头节点序列,从而近似地构建一个翻译推导的层次切分模型,进而来约束翻译推导.

7 结论及后续工作

翻译推导的切分歧义是统计机器翻译中一个很困难的问题,与基于短语的机器翻译相比,由于层次短语翻译模型一些独有的特点,这使得切分歧义问题更加严重.本文通过研究层次短语翻译推导在源语言端的层次切分的歧义问题,提出了一种基于 MRF 的切分模型,这是本文的主要贡献.本文不是直接对规则在源语言端的切分所对应的词序列进行建模,而是对词序列中具有代表意义的少数几个头节点序列建模.这样做可以有效降低模型的训练和解码的复杂度,进而保证了翻译系统的解码速度不受影响;同时,它又不会因为头节点序列所含的信息少、表达能力不足,导致模型没有足够的泛化能力.在 3 个测试集上的实验结果表明,该模型能够有效地提高翻译质量;同时,通过分析实验数据我们发现,获得更好的翻译结果确实是因为本文的切分模型选择了更合理的层次切分.

本文的模型同样适用于基于短语的翻译,我们将在后续的工作中检验它在短语翻译模型中的效果.另外,我们将在切分模型中融入更为丰富的上下文信息,尤其是目标语言端的信息.

References:

[1] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proc. of the HLT-NAACL. Edmonton: Association for Computational Linguistics, 2003. 48−54. [doi: 10.3115/1073445.1073462]

[2] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin a, Herbst E. Moses: Open source toolkit for statistical machine translation. In: Proc. of the ACL on Demonstration Sessions. Prague: Association for Computational Linguistics, 2007. 177−180.

[3] Yamada K, Knight K. A syntax-based statistical translation model. In: Proc. of the ACL. Toulouse: Association for Computational Linguistics, 2001. 523−530. [doi: 10.3115/1073012.1073079]

[4] Liu Y, Liu Q, Lin SX. Tree-to-String alignment template for statistical machine translation. In: Proc. of the ACL-COLING. Sydney: Association for Computational Linguistics, 2006. 609−616. [doi: 10.3115/1220175.1220252]

[5] Chiang D. A hierarchical phrase-based model for SMT. In: Proc. of the ACL. Ann Arbor: Association for Computational Linguistics, 2005. 263−270. [doi: 10.3115/1219840.1219873]

[6] Chiang D. Hierarchical phrase-based translation. Computational Linguistics, 2007,33(2):201−228. [doi: 10.1162/coli.2007.33.2.201]

[7] Bishop CM. Pattern Recognition and Machine Learning. Springer-Verlag, 2006. 383−418.

[8] Collins M. Head driven statistical models for natural language parsing [Ph.D. Thesis]. Pennsylvania: University of Pennsylvania, 1999.

[9] Och FJ. Minimum error rate training in statistical machine translation. In: Proc. of the ACL. Sapporo: Association for Computational Linguistics, 2003. 160−167. [doi: 10.3115/1075096.1075117]

[10] Lin DK. A path-based transfer model for machine translation. In: Proc. of the COLING. Geneva: Association for Computational Linguistics, 2004. [doi: 10.3115/1220355.1220445]

[11] Shen LB, Xu JX, Weischedel R. A new string-to-dependency machine translation algorithm with a target dependency language model. In: Proc. of the ACL. Columbus: Association for Computational Linguistics, 2008. 577−585.

[12] Zollmann A, Venugopal A. Syntax augmented machine translation via chart parsing. In: Proc. of the Workshop on Statistical Machine Translation. New York: Association for Computational Linguistics, 2006. 138−141.

[13] Marton Y, Resnik P. Soft syntactic constraints for hierarchical phrased-based translation. In: Proc. of the ACL. Columbus: Association for Computational Linguistics, 2008. 1003−1011.

[14] Lafferty JD, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of ICML. San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[15] Finkel JR, Kleeman A, Manning CD. Efficient, feature-based, conditional random field parsing. In: Proc. of the ACL. Columbus: Association for Computational Linguistics, 2008. 959−967.

[16] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program, 1989,45(3):503−528.

[17] Wu DK. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 1997, 23(3):377−404.

[18] He ZJ, Liu Q, Lin SX. Improving statistical machine translation using lexicalized rule selection. In: Proc. of the COLING. Manchester: Association for Computational Linguistics, 2008. 321−328.

[19] Liu Q, He ZJ, Liu Y, Lin SX. Maximum entropy based rule selection model for syntax-based statistical machine translation. In: Proc. of the EMNLP. Hawaii: Association for Computational Linguistics, 2008. 89−97.

[20] Cui L, Zhang DD, Li M, Zhou M, Zhao TJ. A joint rule selection model for hierarchical phrase-based translation. In: Proc. of the ACL. Uppsala: Association for Computational Linguistics, 2010. 6−11.

[21] Stolcke A. SRILM—An extensible language modeling toolkit. In: Proc. of the ICSLP. 2002. 901−904.

[22] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Technical Report, TR-10-98, Harvard University, 1998.

[23] Och FJ, Ney H. Improved statistical alignment models. In: Proc. of the ACL. Hong Kong: Association for Computational Linguistics, 2000. 440−447. [doi: 10.3115/1075218.1075274]

[24] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the ACL. Philadelphia: Association for Computational Linguistics, 2002. 311−318. [doi: 10.3115/1073083.1073135]

[25] Doddington G. Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. In: Proc. of the HLT. San Francisco: Association for Computational Linguistics, 2002. 138−145.

[26] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proc. of the ACL. Sapporo: Association for Computational Linguistics, 2003. 423−430. [doi: 10.3115/1075096.1075150]

[27] Bikel DM. On the parameter space of generative lexicalized statistical parsing models [Ph.D. Thesis]. Pennsylvania: University of Pennsylvania, 2004.

[28] Vilar D, Stein D, Ney H. Analysing soft syntax features and heuristics for hierarchical phrase based machine translation. In: Proc. of the IWSLT. 2008.

[29] Chiang D, Knight K, Wang W. 11 001 new features for statistical machine translation. In: Proc. of the HLT-NAACL. Boulder: Association for Computational Linguistics, 2009. 218−226.

[30] Chiang D. Learning to translate with source and target syntax. In: Proc. of the ACL. Uppsala: Association for Computational Linguistics, 2010. 1443−1452.

[31] Xiong DY, Zhang M, Li HZ. Learning translation boundaries for phrase-based decoding. In: Proc. of the NAACL-HLT. Los Angeles: Association for Computational Linguistics, 2010. 136−144.

[32] Stein D, Peitz S, Vilar D, Ney H. A cocktail of deep syntactic features for hierarchical machine translation. In: Proc. of the Conf. on Association for Machine Translation in the Americas (AMTA). 2010.

[33] Gao Y, Koehn P, Birch A. Soft dependency constraints for reordering in hierarchical phrase-based translation. In: Proc. of the EMNLP. Edinburgh: Association for Computational Linguistics, 2011. 1003−1011.

[34] Xiong DY, Zhang M, Aw AT, Li HZ. A source dependency model for statistical machine translation. In: Proc. of the MT Summit XII. 2009.

刘乐茂(1985－),男,江西余干人,博士生,主要研究领域为机器翻译,自然语言处理,机器学习.

赵铁军(1962－),男,博士,教授,博士生导师,主要研究领域为机器翻译,自然语言处理,机器学习.

曹海龙(1976－),男,博士,讲师,主要研究领域为机器翻译,句法分析,自然语言处理,机器学习.

朱聪慧(1979－),男,博士,讲师,主要研究领域为机器翻译,自然语言处理,机器学习.

张春越(1985－),男,博士生,主要研究领域为机器翻译,自然语言处理,机器学习.