软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(9):2347−2357 [doi: 10.3724/SP.J.1001.2012.04165]

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

基于超图的翻译模型融合的研究*

刘宇鹏+, 李 生, 赵铁军

(哈尔滨工业大学 计算机科学与技术系,黑龙江 哈尔滨 150001)

Research on the Translation Model Combination Based on Hypergraph

LIU Yu-Peng+, LI Sheng, ZHAO Tie-Jun

(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

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

Liu YP, Li S, Zhao TJ. Research on the translation model combination based on hypergraph. *Journal of Software*, 2012,23(9):2347−2357 (in Chinese). http://www.jos.org.cn/1000-9825/4165.htm

Abstract: The system combination performs under post-processing, but the paper introduces a translation model combination, which combines two mainstream translation models (Hiero and MaxEnt-based BTG) during decoding. To the spurious ambiguity and consensus problem, the paper introduces new decoding method to solve two problems. In experiment, translation model combination is significantly better than member model, and the new decoding method is better.

Key words: hypergraph; derivation; rule; translation model combination; spurious ambiguity; consensus translation

摘 要: 当前,系统融合是在机器翻译的后处理上进行.提出了在解码过程中来融合翻译模型,融合了主流两个翻译系统的翻译模型(层次化的基于短语的文法 Hiero 和括号转录文法 BTG).并从理论和实践的角度探索了现在主流的两种解码方法.同时,所提出的解码方法解决了伪歧义或一致性问题.在实验结果上得出:多文法模型融合的标志性要好于成员翻译模型;新的解码方法标志性好于传统解码方法(Viterbi 解码).

关键词: 超图;推导;规则;翻译模型融合;伪歧义;一致性翻译

中图法分类号: TP391 **文献标识码**: A

系统融合是把多个系统输出的 *N*-best 结果进行融合,生成新的翻译结果.而且已证明,融合的翻译结果要好于单个系统的输出.按照融合的粒度来分,包括“句子级”、“短语级”、“词级”[1−5].最近,在基于混淆网络的词一级系统融合技术获得的标志的性能,但是这些方法都是在机器翻译的后处理上来进行融合.传统的在后处理上做系统融合的方法没有充分考虑解码过程的信息,而且后处理上的融合不能充分考虑解码中巨大的搜索空间.

模型间的融合在近几年才得到发展,Liu[6]使用超图来完成两个模型的融合,并探讨了两种解码方法,但不是真正意义上的超图,只是在部分翻译中利用超图的思想来完成假设翻译的生成.Li[7]通过一致性特征来完成模型间的互相影响,从而达到融合的目的.这个框架是 *N*-best 上建立的,而不是在表示更大搜索空间超图的基础上.Jiang[8]在推导的过程中更加灵活地运用双语句法和层次化短语的规则,对于两种文法进行合成.Duan[9]提出了

* 基金项目: 国家自然科学基金(60736014); 国家高技术研究发展计划(863)(2006AA010108); 黑龙江省教育厅科学技术研究项目(12521073)

收稿时间: 2011-02-19; 修改时间: 2011-08-31; 定稿时间: 2011-11-02

在翻译模型间进行线性插值来完成模型间的相互影响,两种翻译模型不是可以相互利用的.Duan[10]提出了在翻译模型间进行特征选择,得到一些重要的特征.DeNero[11]是在超图的框架下,通过 *n*-gram 后验概率特征来对两个模型进行重新搜索得到结果.Duan[12]也是在超图框架下,但不同的是通过两个模型的 *n*-gram 后验概率特征进行线性插值得到翻译结果,且采用了两阶段的最小错误率训练.

按照 Liu[6]的思想,翻译模型融合的框架分为两种:翻译级融合和推导级融合;解码框架分为两种:最大翻译解码和最大推导解码.经实验比较,翻译模型融合的框架采用推导级融合,而解码框架采用最大翻译解码效果最好.本文采用了这种框架结构,与 Liu[6]的工作不同的是:

- 我们融合了不同翻译模型是基于最大熵的 BTG[13,14]和层次化短语的 SCFG[15,16],发现两种模型是互补的,融合的模型明显好于单个模型;
- 在解码方法上,Liu[6]仅仅解决伪歧义翻译(最大翻译解码),而没有考虑到解决一致性问题[1−7,23,24]相结合的方法.而伪歧义和一致性是从不同的角度来解决解码的问题,我们生成目标翻译的推导是来自于不同的系统,实验也证明了结合两种方法的有效性;
- 同时,为了证明本文提出方法的有效性,分别在国内和国际知名的机器翻译评测上做了大量实验来验证方法的有效性和稳定性.基线系统选择是国际上认可的两个机器翻译系统,也是本文中提出的融合前的成员模型.

为使两个翻译模型统一在一个框架下,本文第 1 节介绍超图.第 2 节介绍两个单独翻译模型和翻译模型融合方法.第 3 节介绍消除伪歧义和生成一致性翻译的解码,并把两种方法结合在一起.第 4 节是实验结果,分别做了两组实验,使用国内和国际机器翻译方面知名的评测数据集 CWMT09(Chinese Workshop of Machine Translation 2009)和 NIST09(National Institute of Standards and Technology 2009)进行实验以验证结果的有效性.

1 基于超图的翻译模型融合

超图(图的泛化)从 19 世纪 70 年代就开始在离散数学中的许多建模问题上得到了应用.我们也把超图称为有向超图来抽象可以用动态规划来解决的层次化搜索空间,也就是把一个大问题变成子问题分而治之.为了把翻译模型融合到一个框架下,我们引入超图的概念.

1.1 超 图

- 超图:有向超图是一个带有 R 的对 $H=\langle V,E\rangle$,V 是节点的集合,$E\subseteq V^*\times V$ 是超边的集合,R 是权重的集合.
- 超边:每一个超边 $e\in E$ 是一个三元组 $e=\langle T(e),h(e),f_e\rangle$,其中,$T(e)\in V^*$是尾节点的有序序列,$h(e)\in V$ 是头节点序列,f_e 是一个从 $R|T(e)|$到 R 的权函数.
- 超节点:与超边相关联的尾节点有序序列 $T(e)$和头节点序列 $h(e)$都称为超节点,每个头超节点都与多个超边相连.
- 元数:我们定义$|e|=|T(e)|$是超边的元数.如果$|e|=0$,那么 $f_e\in R$ 是一个常量(f_e 是一个空函数).同时,我们把 $h(e)$称为源节点.超图中所有超边的最大元数为超图的元数.元数为 1 的超边是正则边,元数为 1 的超图为正则图(格).词图(word lattice)就是元数为 1 的超图.
- 翻译超图:建立在超图的基础上,一个翻译规则对应一条超边(推导);翻译规则的权重对应超边的权函数.翻译节点是在翻译过程中生成的部分翻译,且带有各种特征值.

1.2 成员翻译模型

由于本文将 SCFG 和 BTG 引入统计机器翻译,为了更好地说明翻译模型的融合,我们首先给出单个翻译模型具体的概念.

1. 同步上下文无关文法(SCFG)是一个五元组 $G=\langle \Sigma_s,\Sigma_t,N,R,S\rangle$:

- Σ_s 和Σ_t 分别是源语言端和目的语言端的终结符(词语、单词)字符集;
- N:非终结符集合,$N\cap(\Sigma_s\cup\Sigma_t)\Sigma_t=\varnothing$;

- $R\subset N\times(\Sigma_s\cup(N\times Num))^*\times(\Sigma_t\cup(N\times Num))^*$,$Num$ 是数字的集合.对于每一个规则$(A,u,v),(B,n)\in N\times Num$ 既出现在源语言片断中,也应该出现目标语言片断中.我们把规则(A,u,v)表示成为 $A\rightarrow\langle u,v\rangle$,把$(B,n)$表示为 B_n.

作为示例,我们定义一个 SCFG 的五元组 $G=\langle\Sigma_s,\Sigma_t,N,R,S\rangle$.初始的非终结符 S 的规则(粘合规则)可以表示为

$$S\rightarrow\langle SX,SX\rangle \tag{1}$$

$$S\rightarrow\langle X,X\rangle \tag{2}$$

本文采用的成员翻译模型 1 为基于 SCFG 的 Hiero[15,16].层次短语规则可以表示为

$$X\rightarrow\langle yu\ X_1\ you\ X_2,have\ X_2\ with\ X_1\rangle \tag{3}$$

$$X\rightarrow\langle X_1\ zhiyi,one\ of\ X_1\rangle \tag{4}$$

传统的短语规则可以表示为

$$X\rightarrow\langle beihan,North\ Korea\rangle \tag{5}$$

2. BTG(括号转录文法):ITG(反转转录文法)的一种退化形式,BTG 只有一个非终结符,其规则也只有下面 3 种:

$$X\xrightarrow{[\cdot]}(X^1,X^2) \tag{6}$$

$$X\xrightarrow{\langle\cdot\rangle}\langle X^1,X^2\rangle \tag{7}$$

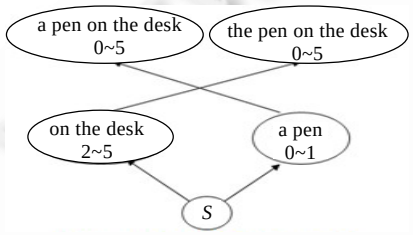
$$X\rightarrow(u,v) \tag{8}$$

规则(6)用于保序地合并两个相邻成分,规则(7)用于逆序地合并两个相邻成分,规则(8)用于翻译源语言的单词/短语.本文所采用的成员翻译模型为基于 BTG 的最大熵模型[13,14].

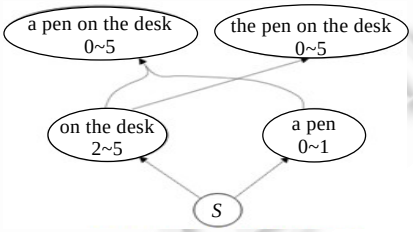
1.3 翻译模型融合策略

在翻译模型融合策略中,不同翻译模型共享超图的同一个节点.例如,虽然在翻译规则中 SCFG 和 BTG 的源语言不同,但是可以翻译源语言的同一个跨度(span);在目标语言侧,BTG 和 SCFG 都生成一个目标语言.

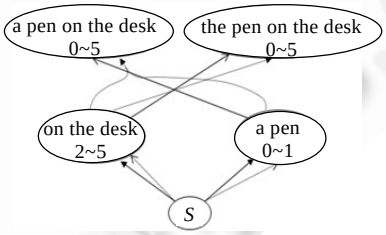
为了说明这种融合方法,如图 1 所示,虚线是 BTG 的规则,实线是 SCFG 的规则.



(a) 由基于层次短语的 SCFG 生成的翻译超图



(b) 由基于最大熵的 BTG 生成的翻译超图



(c) 两种翻译模型融合超图的压缩形式

Fig.1 An example of packed hypergraph of translation model combination

图 1 翻译模型融合的翻译超图实例

如果要翻译源语言“*zhuozi shang gangbi*”,对于图 1(a),可能匹配上 SCFG 规则:

$$X\rightarrow\langle zuozi\ shang\ X_1,X_1\ on\ the\ desk\rangle \tag{9}$$

$$X\rightarrow\langle X_1\ gangbi,open\ X_1\rangle \tag{10}$$

匹配上的词汇化规则:

$$X\rightarrow\langle gangbi,the\ pen\rangle \tag{11}$$

$$X\rightarrow\langle zuozi\ shang,on\ the\ desk\rangle \tag{12}$$

对于图 1(b),可能匹配上 BTG 规则:

$$X \xrightarrow{\langle\cdot\rangle} zuo\ zi\ shang / on\ the\ desk, gangbi / a\ pen \tag{13}$$

在图 1(c)中发现,由于 SCFG 和 BTG 都生成“*a pen on the desk*”和“*the pen on the desk*”,可以共享上面的 SCFG 和 BTG 翻译节点(部分翻译).

2 基于翻译模型融合数学模型

机器翻译系统的基本任务是将一个输入的源语言句子 f 翻译成适当的目标语言句子 e.这样,机器翻译模型的基本任务就是对概率 $P(e|f)$进行建模.本模型是通过超图来对翻译过程进行建模.因此,首先需要引入隐含变量推导 d 来表示 SCFG 和 BTG 的规则.于是,$P(e|f)$可以得到如下推导:

$$p(e\mid f)=\sum_{d}p(e,d\mid f) \tag{14}$$

接着,我们对公式(14)右边和式中的每项进行推导:

$$P(e,d|f)=P(d|f)P(e|d,f) \tag{15}$$

公式(15)将公式(14)中等号右边和式中的每项分解为 2 个因式,每个因式对应一个子模型.其中,第 2 个子模型 $p(e|d,f)$对应从源语言和推导中得到目标语言的过程.对于一个既定的源语言句子和规则,有且仅有 1 个目标语言句子与之对应,所以这个子模型可以不予考虑.本研究最关键的是第 1 个子模型 $p(d|f)$,它已知源语言怎么估计每个推导的概率.接下来重点讨论如何基于翻译模型融合地对这个子模型进行建模:

$$p(e,d\mid f)=p(d\mid f)=\frac{e^{\gamma h(e,f,d)}}{\sum_{\gamma,d}e^{\gamma h(e,f,d)}} \tag{16}$$

其中,$h(e,f,d)$是特征向量,γ是特征权重向量.

在本文实现的系统中,采用了以下特征:

- 双向翻译概率:$\phi(e|f)$和$\phi(f|e)$;
- 双向词汇化翻译概率:$lex(e|f)$和 $lex(f|e)$;
- 语言模型得分:$lm(e)$;
- 翻译过程用到规则的个数:$Num(Rule)$;
- 生成翻译的单词数:$Numword(Sent)$;
- 翻译过程中用到短语的个数:$Numphrase(Sent)$;
- 基于最大熵的扭曲模型的概率:$D(e,f)$.

在翻译模型融合中用到这 9 个特征,而两个成员模型仅仅用到了 8 个特征.在 SCFG 模型中,使用了除了基于最大熵的扭曲模型的概率 $D(e,f)$的其他的 8 个特征;在 BTG 模型中,使用了除了翻译过程用到规则的个数 $Num(Rule)$的其他的 8 个特征.两个模型共用了除了 $D(e,f)$和 $Num(Rule)$的 7 个特征,其实,这两个特征都反映了在推导过程中使用规则的情况,这些能够保证两个模型的一致特征.所以,两个模型的 8 个特征具有可比性,融合后的 9 个特征能够反映两个成员模型的融合情况.

3 解码方法

因为在统计机器翻译奠基性的工作中,利用信源信道模型对翻译过程进行建模,所以在随后统计机器翻译的研究中,都约定俗成地将翻译过程称为解码.一个机器翻译系统的任务就是把输入的源语言句子 f 翻译成为目的语言句子 e.也就是说,一个机器翻译过程就是在所有目的语言候选$\{e\}$中找出最优的译文的过程.翻译过程

可以分为 5 个步骤:

第 1 步:加载翻译模型(文法规则集)和语言模型;

第 2 步:读入源语言句子 *f*,生成不同的源语言片段;

第 3 步:获取可用的规则集;

第 4 步:进行从底向上的源语言到目标转化过程,这个过程称为栈搜索;并进行两种解码方法的结合:一种是消除伪歧义,另一种是消除一致性;

第 5 步:将最优译文输出.

为了把两个解码器进行融合,要求解码的样式(生成目标语言的顺序和解码的次数)是一样的.解码算法按照生成目标语言的顺序:1) 从左到右[18];2) 自底向上[13−16].按照解码的次数为:一阶段解码[13−16,18];多阶段解码[19−21].

按照解码时生成假设的情况有两类:

1. 伪歧义解码[6,17]

即一个输出串的概率有许多不同的推导(例:部分树结构或是部分分词结果).原则上,翻译结果的好坏应该由生成它的推导总概率值来表示.然而,找到最好的翻译(解码)是计算不可行的.因此,大部分系统使用 Viterbi 近似生成一个好的翻译结果(并进行了一定的剪枝).

2. 一致性解码

所谓的一致性解码是为了生成的最好结果与其余 *N*-best 结果一致;按照生成一致性翻译系统的个数分为:系统间的一致性解码[1−7]和系统内的一致性解码[23,24].

整个解码器是建立在 CYK 算法基础上的.为了避免搜索所有的推导,我们采用了 beam 搜索策略,也就是在搜索过程中将一些不好的推导删除掉.整个核心算法见算法 1 和算法 2.对于超图,同一个节点中有不同 SCFG 或是 BTG 推导,为了更好地利用两种规则.我们首先从消除伪歧义角度对解码公式进行推导:

$$\begin{aligned}\tilde{e} &= \arg\max(P(f\mid e))\\ &= \arg\max\left(\sum_{d\in D(e,f)} P(f,d\mid e)\right)\\ &\approx \arg\max\left(\max_{d\in D(e,f)}(P(f,d\mid e))\right)\\ &\approx \arg\max\left(\sum_{d\in D(e,f)\cap ND(e)} P(f,d\mid e)\right)\end{aligned} \tag{17}$$

接着从生成一致性翻译来对解码公式进行推导:

$$\begin{aligned}\tilde{e} &= \arg\min RISK(e)\\ &= \arg\min \sum_{e'\in R(f)} LOSS(e,e')P(e'\mid f)\\ &\approx \arg\min \sum_{e'\in T'(f)} LOSS(e,e')P(e'\mid f)\end{aligned} \tag{18}$$

为了结合两种方法,对上面的两种解码方法进行说明:$p(e|f)$表示由源语言生成目标语言的概率,$p(e,d|f)$为加入隐变量推导后的概率,D 表示关于原语言和目标语言所有推导的集合,$ND(e)$表示生成 *N*-best 翻译结果,$RISK(e)$是生成目标语言的贝叶斯风险函数,$LOSS(e,e')$是为了计算最小贝叶斯风险的一个损失函数,$R(f)$是参考假设空间,$T'(f)$是验证假设空间.为了消除伪歧义,我们把一致性解码中的源语言到目标语言的翻译概率变成生成同样目标语言的和,即

$$p(e'\mid f)=\sum_{d\in D} p(e',d\mid f) \tag{19}$$

这样,不同于传统的 Viterbi[18]方式解码,既能解决解码中的一致性问题,又能消除伪歧义.这种方式对于翻译模型的融合更加重要,因为有很多的推导来自于不同的翻译模型.

算法 1. *addEdge*(*e*).

Threshold pruning

```
1:  if e.score<bestCurrentScore[e.start,e.end]−beam then
2:    discard e and return
3:  end if
    /*select decoding strategy*/
4:  select case:
5:    Viterbi:
6:      if there is a matching edge e* in chart[e.start,e.end] then
7:        if e.score>e*.score then
8:          replace e* with e and return
9:        else
10:         discard e and return
11:       end if
12:     end if
13:   Crunching:
14:     if there is a matching edge e* in chart[e.start,e.end] then
15:       e.score=e.score+e*.score
16:     end if
17: end select

    /*Historgram pruning*/
18: if |chart[e.start,e.end]|>b then
19: select case:
20:   Viterbi:
21:     if e.score>e*.score then   /*e* is the worst edge in chart[e.start,e.end]*/
22:       replace e* with e and return
23:     else
24:       discard e and return
25:     end if
26:   Crunching:
27:     if e.target==e*.target then   /*e* is the n-best in in chart[e.start,e.end]*/
28:       e.score=e.score+e*.score
29:     end if
30: end select
31: end if
```

为了消除伪歧义问题,我们引入算法 1.对于同一部分翻译,我们采用 Viterbi 和 Crunching 两种解码方法:

- 如果是 Viterbi,我们选取最高得分;
- 如果是 Crunching,我们把部分特征得分加和,当然,语言模型和词惩罚等公用特征得分不需要加和.

算法 2. *joint_decoder*(*sentence s*).

```
    //Load BTG translation model
1:  LoadTranslationModel(BTG);
    //Load SCFG translation model
2:  LoadTranslationModel(SCFG);
```

```
    /*for all lexical rules with the matching source side on the consecutive sequence of s do Generate a new
      edge e(addEdge(e))*/
3:  initializeChartItem(s);
4:  for span=2 to n do
5:    for start=1 to n-span+1 do
6:      end=start+span−1
       /*For each middle position, combine edge[start,middle] with edge[middle,end], using BTG/SCFG rule
        to generate a new edge[start,end]*/
7:      addEdge(e)
8:      CompleteChartItem(start,end)
        /*use loss function regenerate the overall score*/
9:      If decoding style is MBR then
10:       SortHypoUseByLoss(start,end)
11:     end if
12:   end for
13: end for
```

翻译模型融合解码的算法(算法 2)采用 CYK 算法,是一个很直观的过程.根据融合翻译模型的个数来加载翻译模型;利用成员翻译模型,对于每个源语言片段生成 *N*-best 翻译;如果采用了 MBR 解码,利用损失函数和生成部分翻译的得分来对生成的 *N*-best 翻译重新排序,并得到一个新的得分.

4 实 验

本节给出了基于 SCFG 的翻译模型与基于 BTG 翻译模型融合(简称为 TMCOM)以及基于单个模型的实验对比分析.对于单个成员模型,我们采用由 Chiang 等人开发的 Hiero 系统[15,16]和熊德意的 MaxEnt_based BTG 系统[13,14].两个系统都是影响很大的基于形式句法的统计机器翻译系统.以下实验在 CWMT09 和 NIST09 评测的数据集上进行.因为本文所提出的模型是一种融合的翻译模型,即在训练双语语料的源语言端和目的语言端抽出短语表后,还要生成 SCFG 和 BTG 规则.为了验证栈的大小对于翻译性能的影响,我们测试了不同栈空间对翻译性能带来的影响.

4.1 实验1:基于CWMT09语料的实验

4.1.1 实验设置

本组实验所用语料是 CWMT09 评测所有语料一部分.训练集共 3 804 000 句对(数据包括 CLDC-LAC-2003-2006、万方数据、点通数据、厦门大学电影字幕、哈尔滨工业大学信息检索数据等).语言模型用的双语语料英文部分和路透社的 RCV1 部分.开发集是从几年的 CWMT 评测集合中随机选出来共 503 句.测试集是 CWMT07 共 1 002 句.本数据集中对每个中文句子,有 4 个英文参考译文与之对应.

本实验中采用了 GIZA++词对齐工具[22]来获取训练句对的词对齐信息.基于语言模型的英文句子,我们用 SRILM 工具[25]训练了一个五元的语言模型,采用的平滑算法为修正的 Kneser-Ney 策略[26].实验所用各模型的特征权重均采用文献[27]中提出的最小错误率训练方法进行估计.此方法在开发集上针对评测指标 BLEU (bilingual evaluation understudy)[28]进行最优化迭代,来对参数权重进行调整.

本文采用 BLEU 作为译文质量的评价指标,并用 NIST(National Institute of Standards and Technology)官方网站发布的 mteval-v12.pl 来进行计算.实验结果的 BLEU 分数采用自举重抽样方法(bootstrapping resampling)[29]进行显著性测试,测试工具为 Zhang 等人的实现[30].如果不作特殊说明,则以下所报结果均具有 95%的置信度.

4.1.2 系统设置

本实验采用的作为对比的基准系统是 Heiro[15,16]和 BTG[13,14],它是一个广泛流行的基于短语的机器翻译系统.Heiro 采用了 8 个默认的特征,而 BTG 采用了 8 个特征(基于最大熵的扭曲模型的概率:$D(e,f)$).在抽出的短语表中,源语言的最大长度为 5,而目标语言的最大长度为 3.对 SCFG 系统,双向短语的最大词长为 10,规则中抽象非终结符个数的上限设置为 2.不允许源语言侧两个非终结符相邻.解码时的 BeamThreshold 设置为 30.而在 BTG 中,使用最大熵工具包中[31]的 GIS[32]算法来建立扭曲模型.

4.1.3 实验结果

表 1 给出了 3 个系统在开发集上用最小错误率训练估计出的特征权重值.SCFG/TMCOM 系统均没有词序扭曲模型,其他特征均可类比于 Pharaoh 的特征.比如:SCFG 系统的规则概率类似于 Pharaoh 中的短语翻译概率,SCFG 系统的词汇互译得分特征类似于 Pharaoh 的词汇化权重,SCFG 系统的规则数类似于 Pharaoh 的短语个数惩罚 pp,译文词个数惩罚与语言模型得分特征 3 个系统一致.BTG 除了没有规则个数惩罚且加入了扭曲模型特征外,与 Pharaoh 的特征相同.比较表 1 中的各个系统权重可以看出,3 个系统的语言模型权重均比较大,表明语言模型是一个很重要的特征.

Table 1 Feature weights of combination and individual model obtained by MER training on the development set

表 1 翻译模型融合和成员模型在开发集上通过最小错误率训练获得的特征权重

System	$P(c\|e)$	$P(e\|c)$	$P_w(c\|e)$	$P_w(e\|c)$	$P_{lm}(e)$	$P_r(e)$	Word penalty	Phr. penalty	Rule penalty
Hiero	1.578 5	0.347 8	0.558 1	1.613 9	3.570 1	—	3.946 3	−0.453 8	−2.989 0
BTG	0.897 1	0.136 3	0.466 6	0.332 9	0.495 2	2.960 7	2.960 7	−1.515 7	—
TMCOM	1.383 1	0.389 8	0.473 9	1.539 2	2.953 4	6.230 0	3.284 1	0.287 6	−2.418 8

表 2 给出了 3 个系统的性能比较,从中可以看出:

1) TMCOM 系统显著地超过了 BTG 和 Hiero.在 BLEU-4 分值上,TMCOM 系统比 BTG 有 0.0116(0.2659−0.2543)的绝对提升,即相对性能提高为 4.56%(0.0116/0.2543);比 Hiero 有 0.0070(0.2659−0.2567)的绝对提升,即相对性能提高为 2.72%(0.0070/0.2567).这些结果表明,基于 SCFG 和 BTG 的机器翻译模型只能有效地对不同的调序结构进行建模.本文提出的 TMCOM 的模型能够有效地捕捉两种文法互补的部分;
2) 同时,Crunching-MBR 显示出 Viterbi 解码的绝对优势.这是因为对于生成相同的部分翻译,Crunching 增加了特征值,而 Viterbi 仅仅选取最高得分;同时,MBR 也完成了部分翻译间的一致性验证,使得与其他部分翻译最相似的部分翻译得到选择;
3) 同时我们还可以看出,BTG 系统的性能也劣于 Hiero.主要原因在于,BTG 仅仅完成相邻源语言/目标语言的调序,而 Hiero 获得的是不连续源语言/目标语言的调序,这一点上两种文法也存在互补性.

Table 2 System performances

表 2 系统性能

Model	CWMT07
TMCOM-Crunching-MBR	0.2659±0.0073
TMCOM-Viterbi	0.2608±0.0089
Hiero	0.2567±0.0101
BTG	0.2543±0.0089

图 2 给出了各个系统在测试集上不同栈大小的性能比较.可以看到,随着栈的扩大,系统的性能是逐渐提高的.其中,横坐标为 1,2,3,4,5 对应栈的规模分别是 10,30,50,100,200,而且基于超图的翻译模型融合都超过了成员模型.

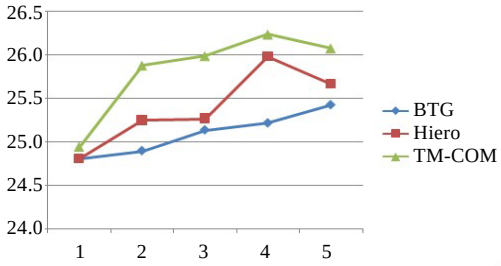


Fig.2 Translation result of translation model combination at the condition of different stack

图 2 翻译模型融合在不同栈大小时翻译结果

4.2 实验2:基于NIST09评测语料的实验

4.2.1 实验设置

本实验在2009年NIST机器翻译评测中的中英翻译任务上进行.抽取规则集所用的训练句对来自语言学数据联盟(linguistic data consortium)提供的MTC数据集(编号:LDC2002T01)、中文树库双语语料(编号:LDC2003E07)、FBIS数据集(编号:LDC2003E14)、UN中英文双语语料(编号:LDC2004E12)、中文新闻翻译语料(编号LDC2005T06).该数据集经过过滤,包含约4 676 000句对.语言模型的训练语料是GIGAWORD语料中的新华部分和双语语料英文部分.语言模型采用SRILM工具进行训练,设置元数为五元.

另外,在本文的实验中,先对训练数据进行两个方向的GIZA++[22]词对齐,然后采用grow-diag-final启发式规则获得多对多的词对齐关系.其他实验设置与实验1相同.评测准则本组实验利用了BLEU和NIST两种指标.其他实验设置与实验1相同.系统设置和实验1一样.

4.2.2 实验结果

表3和表4分别给出了3个系统在实验2数据集上的实验BLEU分值对比和NIST分值对比情况.

Table 3 Comparison BLEU results of individual translation model and translation model combination on 2005 NIST Chinese-English test set

表 3 在2005年NIST中英机器翻译评测集上单个翻译模型和翻译模型融合的对比结果(BLEU)

System	BLEU-*n*	*n*-Gram precision								
	4	1	2	3	4	5	6	7	8	9
TMCOM	0.262 4	0.738 3	0.366 9	0.183 6	0.095 4	0.051 4	0.028 3	0.016 4	0.010 0	0.006 1
Hiero	0.246 5	0.728 9	0.347 7	0.170 7	0.085 3	0.042 3	0.020 2	0.009 9	0.005 0	0.002 4
BTG	0.241 8	0.741 1	0.348 7	0.165 2	0.080 0	0.039 8	0.020 1	0.010 5	0.005 5	0.002 8

Table 4 Comparison NIST results of individual translation model and translation model combination on 2005 NIST Chinese-English test set

表 4 在2005年NIST中英机器翻译评测集上单个翻译模型和翻译模型融合的对比结果(NIST)

System	NIST-*n*	*n*-Gram precision								
	4	1	2	3	4	5	6	7	8	9
TMCOM	8.346 8	6.370 3	1.572 8	0.321 5	0.064 8	0.017 4	0.005 9	0.002 2	0.000 8	0.000 2
Hiero	8.217 9	6.365 9	1.488 5	0.292 9	0.056 2	0.014 4	0.004 5	0.001 3	0.000 4	0.000 1
BTG	8.132 5	6.265 3	1.480 8	0.310 1	0.060 6	0.015 8	0.005 1	0.001 5	0.000 7	0.000 2

从表3和表4中我们可以看出,对比结果与实验1结果一致,基于使用新的解码方法的翻译模型融合的系统仍然稳定地超过单个翻译模型.与实验1相对照,我们还可得出以下分析结论:

1) 实验1所用数据集为CWMT09年的训练语料,测试语料是CWMT07的测试语料;而实验2所用数据集为NIST09的训练语料,测试语料为NIST05.综合两组实验说明,新提出的翻译模型融合在两种评测语料的数据集上都能够取得稳定的性能优势;

2) 实验1和实验2所用基线系统是翻译模型融合的成员翻译模型;而实验2在每一个*n*-gram精度上都

超过了单个翻译模型.

另外,通过仔细分析表 3 中的 *n*-gram 精确度我们发现:在 1-gram~4-gram 的精确度上,使用新的解码方法的翻译模型融合的系统均明显高于单个系统;但是在 5-gram~9-gram 的精确度上,使用新的解码方法的翻译模型融合的系统优势不大.BLEU 指标的一个特性就是它偏向于检测语言的流利度,而流利度主要由较长的词串体现.这个细节表明,使用新的解码方法翻译模型融合的系统在得到流利度更高的句子上效果不是很好,这在本质上也与长距离调序以及非连续短语模拟能力相关,说明 Hiero 和 BTG 在生成流利度更高的句子存在缺陷,如果能把目标语言是句法的系统集成到该翻译模型融合框架下,相信能得到更好的结果.

5 结 论

本文提出一种基于超图的统计机器翻译模型融合,这一模型可以在解码阶段对于不同文法结构的翻译模型进行建模.本文给出了基于翻译模型融合解码算法,并且解决了解码中存在的伪歧义和一致性问题.对于解码问题,本文提出一种自底向上的、翻译节点依次扩展的集束搜索算法.在两组国内和国际评测的数据集上进行的实验,均验证了基于翻译模型融合相对于基于 PCFG 和基于 BTG 模型的稳定优势.在后续的研究中,我们将深入研究把基于句法的翻译模型融合到该方法中.

References:

[1] Rosti AVI, Ayan NF, Xiang B, Matsoukas S, Schwartz R, Dorr B. Combining outputs from multiple machine translation systems. In: Proc. of the HLT/NAACL. New York: Association for Computational Linguistics. 2007. 228−235.

[2] He XD, Yang M, Gao JF, Nguyen P, Moore R. Indirect-HMMbased hypothesis alignment for combining outputs from machine translation systems. In: Proc. of the EMNLP 2008. Honolulu: Association for Computational Linguistics. 2008. 98−107.

[3] Rosti AV, Matsoukas S, Schwartz R. Improved word-level system combination for machine translation. In: Proc. of the ACL 2007. Prague: Association for Computational Linguistics.2007. 312−319.

[4] Li CH, He XD, Liu YP, Xi N. Incremental HMM alignment for MT system combination. In: Proc. of the ACL 2009. Singapore: Association for Computational Linguistics. 2009. 949−957.

[5] Chen BX, Zhang M, Li HZ, Aw A. A comparative study of hypothesis alignment and its improvement for machine translation system combination .In: Proc. of the ACL 2009. Singapore: Association for Computational Linguistics, 2009. 941−948.

[6] Liu YP, Mi H, Feng Y, Liu Q. Joint decoding with multiple translation models. In: Proc. of the ACL 2009. Singapore: Association for Computational Linguistics, 2009. 576−586.

[7] Li M, Duan N, Zhang D, Li CH, Zhou M. Collaborative decoding: partial hypothesis re-ranking using translation consensus between decoders. In: Proc. of the ACL 2009. Singapore: Association for Computational Linguistics, 2009. 585−592.

[8] Jiang HF, Yang MY, Zhao TJ, Li S, Wang B. A statistical machine translation model based on a synthetic synchronous grammar. In: Proc. of the ACL 2009. Singapore: Association for Computational Linguistics, 2009. 125−128.

[9] Duan N, Li M, Xiao T, Zhou M. The feature subspace method for SMT system combination. In: Proc. of the EMNLP 2009. Honolulu: Association for Computational Linguistics, 2009. 1096−1104.

[10] Duan N, Li M, Zhang DD, Zhou M. Mixture model-based minimum Bayes risk decoding using multiple machine translation systems. In: Proc. of the COLING 2010. Beijing: Coling 2010 Organizing Committee. 2010. 313−321.

[11] DeNero J, Kumar S, Chelba C, Och F. Model combination for machine translation. In: Proc. of the NAACL 2010. Los Angeles: Association for Computational Linguistics. 2010. 975−983.

[12] Duan N, Li M, Zhang DD, Zhou M. Hypothesis mixture decoding for statistical machine translation. In: Proc. of the ACL 2011. Portland: Association for Computational Linguistics. 2011. 1258−1267.

[13] Xiong DY, Liu Q, Lin SX. Maximum entropy based phrase reordering model for statistical machine translation. In: Proc. of the ACL-COLING 2006. Sydney: Association for Computational Linguistics. 2006. 521−528. [doi: 10.3115/1220175.1220241]

[14] Xiong DY, Zhang M, Aw A, Li HZ. Linguistically annotated BTG for statistical machine translation. In: Proc. of the COLING 2008. Manchester: Coling 2008 Organizing Committee. 2008. 1009−1016.

[15] Chiang D. A hierarchical phrase-based model for statistical machine translation. In: Proc. of the ACL 2005. Ann Arbor: Association for Computational Linguistics. 2005. 263−270.

[16] Chiang D. Hierarchical phrase-based translation. Computational Linguistics, 2007,33(2):1100−1123. [doi: 10.1162/coli.2007.33.2.201]

[17] Li Z, Eisner J, Khudanpur S. Variational decoding for statistical machine translation. In: Proc. of the ACL 2009. Sapporo: Association for Computational Linguistics. 2009. 48−54.

[18] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proc. of the NAACL 2003. Suntec: Association for Computational Linguistics. 2003. 593−601. [doi: 10.3115/1073445.1073462]

[19] Huang L, Chiang D. Forest rescoring: Faster decoding with integrated language models. In: Proc. of the ACL 2007. Prague: Association for Computational Linguistics. 2007. 144−151.

[20] Huang L. Forest reranking: Discriminative parsing with non-local features. In: Proc. of the ACL 2008. Columbus: Association for Computational Linguistics. 2008. 586−594.

[21] Zhang H, Gildea D. Efficient multipass decoding for synchronous context free grammars. In: Pro. of the ACL 2008. Columbus: Association for Computational Linguistics. 2008. 209−217.

[22] Tromble R, Kumar S, Och FJ, Macherey W. Lattice minimum bayes risk decoding for statistical machine translation. In: Proc. of the EMNLP 2008. Honolulu: Association for Computational Linguistics. 2008. 620−629.

[23] Kumar S, Byrne W. Minimum Bayes-risk decoding for statistical machine translation. In: Proc. of the HLT-NAACL 2004. 2004.

[24] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003,29(1):19−51. [doi: 10.1162/089120103321337421]

[25] Stolcke A. SRILM: An extensible language modeling toolkit. In: Proc. of the ICSLP 2002. 2002.

[26] Kneser R, Ney H. Improved backing-off for *M*-gram language modeling. In: Proc. of the ICASSP 2001. 2001. 533−536. [doi: 10.1109/ICASSP.1995.479394]

[27] Och FJ. Minimum error rate training in statistical machine translation. In: Proc. of the ACL 2003. Sapporo: Association for Computational Linguistics. 2003. 160−167. [doi: 10. 3115/1075096.1075117]

[28] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the ACL 2001. 2001. 311−318. [doi: 10.3115/1073083.1073135]

[29] Koehn P. Statistical significance tests for machine translation evaluation. In: Proc. of the EMNLP 2004. Barcelona: Association for Computational Linguistics. 2004. .[doi: 10.1007/ s10590-010-9073-6]

[30] Zhang Y, Vogel S, Waibel A. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In: Proc. of the LREC 2004. Lisbon: Association for Computational Linguistics. 2004. 2051−2054.

[31] Zhang L. Maximum entropy modeling toolkit for python and C++. 2004. http://homepages.inf.ed.ac.uk/s0450736/maxent toolkit.html

[32] Berger A, Pietra SD, Pietra VD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1):39−71.

刘宇鹏(1978－),男,黑龙江哈尔滨人,博士,讲师,主要研究领域为自然语言处理,机器翻译.

李生(1943－),男,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.

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